# PENGARUH INFORMATION GAIN DAN NORMALISASI KATA PADA ANALISIS SENTIMEN BERBASIS ASPEK

Reza Lutfi Nurdiansyah[1], Kania Evita Dewi[2]

[1,2] Teknik Informatika Universitas Komputer Indonesia
Jl. Dipati Ukur No. 112-116, Lebakgede, Kecamatan Coblong, Kota Bandung, Jawa barat 40132
E-mail : kania.evita.dewi@email.unikom.ac.id[2]

**Abstrak**

Informasi di internet sangat beragam, namun begitu banyak pendapat yang mempersulit pengguna lain untuk mendapatkan informasi. Analisis sentimen adalah proses menganalisis atau mengidentifikasi pendapat seseorang terhadap subjek atau produk tertentu yang termasuk dalam kategori positif, negatif, atau netral. Analisis sentimen tingkat aspek menunjukkan kinerja yang lebih baik dibandingkan level dokumen dan level kalimat. Tujuan dari penelitian ini yaitu untuk mengetahui performansi akurasi dari optimalisasi fitur menggunakan *Information Gain* dengan normalisasi kata pada analisis sentimen berbasis aspek. Hasilnya, Support Vector Machine dan perbaikan kata non-standar dengan kamus digunakan dalam penelitian ini sebagai algoritme klasifikasi dengan kernel polinomial juga perbaikan kata tidak baku menggunakan kamus *Slang Word* dan Singkatan (SS) dilanjutkan *Spelling Corrector* menggunakan algoritma *Peter Norvig* dengan tambahan seleksi fitur *Information Gain* untuk mengoptimalkan jumlah fitur. Data yang digunakan adalah sebanyak 1020 data dengan *K-fold* sebesar 10. Berdasarkan hasil pengujian yang telah dilakukan dengan menggunakan *K-fold Cross Validation* dan *Confusion Matrix* terhadap data uji mendapatkan hasil akurasi yang berbeda-beda sesuai dengan alur proses pengujian. Akurasi terbaik didapatkan dari penggunaan *Information Gain* tanpa normaliasasi kata *Peter Norvig* menghasilkan rata-rata akurasi sebesar 83%. Kesalahan yang sering ditemukan ialah pada saat pengubahan kata. Terjadinya kesalahan ini yaitu dikarenakan kata yang dapat diubah hanya dapat mengoreksi satu huruf yang salah. Sehingga perlu dicari lagi metode untuk mengkoreksi kata yang lebih baik.

**Kata kunci:** Analisis Sentimen Berbasis Aspek, Klasifikasi, *Support Vector Machine*, *Information Gain*, Normalisasi, *Peter Norvig*

***Abstract***

*Information on the internet is very diverse, yet so many opinions make it difficult for other users to get information. Sentiment analysis is the process of analyzing or identifying a person's opinion on a particular subject or product that falls into positive, negative, or neutral categories. Aspect-level sentiment analysis shows better performance than document-level and sentence-level. This study aims to determine the accuracy performance of feature optimization using Information Gain with word normalization in aspect-based sentiment analysis. Therefore, this research uses Support Vector Machine as a classification algorithm with a polynomial kernel as well as non-standard word repair using Slang Word and Abbreviation (SS) dictionary followed by Spelling Corrector using Peter Norvig algorithm with additional Information Gain feature selection to optimize the number of features. Based on the test results that have been carried out using K-fold Cross Validation and Confusion Matrix on test data, the accuracy results vary according to the testing process flow. The best accuracy is obtained from the use of Information Gain without Peter Norvig's word normalization resulting in an average accuracy of 83%. Errors are often found when changing words. This error occurs because the word that can be changed can only correct one wrong letter. So in future research, better methods need to be used to correct words.*

***Keywords:*** *Aspect-Based Sentiment Analysis, Classification, Support Vector Machine, Information Gain, Normalization, Peter Norvig*

## 1. PENDAHULUAN

Informasi di internet sangat beragam, salah satunya adalah informasi tentang perspektif pelanggan dari sudut pandang wisata pantai Malang Selatan. Pendapat pelanggan dapat diperoleh dalam bentuk ulasan di platform online. Namun, banyak pendapat yang mempersulit pengguna lain untuk mendapatkan informasi

ini. Analisis sentimen adalah proses menganalisis atau mengidentifikasi pendapat seseorang yang menunjukkan sikap terhadap subjek atau produk tertentu yang termasuk dalam kategori positif, negatif, atau netral [1]. Secara umum, analisis opini dibagi menjadi tiga level, yaitu level dokumen, level kalimat, dan level aspek. Penelitian ini menggunakan analisis sentimen pada tingkat aspek, analisis sentimen level aspek menunjukkan kinerja yang lebih baik dibandingkan pada level dokumen dan kalimat. Ini karena ketika Anda mengungkapkan pendapat Anda cenderung membahas setiap aspek suatu entitas, bukan keseluruhan [2]. Dalam analisis sentimen berbasis aspek, aspek adalah komponen atau atribut dari suatu entitas. Produk, layanan, topik, pertanyaan, orang, organisasi, atau peristiwa dapat dianggap sebagai entitas [3].

Analisis sentimen membutuhkan beberapa hal yang harus dipersiapkan sebelumnya, salah satunya adalah memilih *classifier* mana yang akan digunakan. Metode klasifikasi yang memungkinkan data dibagi menjadi beberapa kategori [4]. *Support Vector Machine* (SVM) dipilih sebagai *classifier* dalam penelitian ini, yang merupakan salah satu metode klasifikasi dan regresi yang paling populer, efisien dan akurat [5]. Namun proses klasifikasi *Machine Learning* seringkali menemukan kelemahan dalam pemilihan fitur atau parameter yang dapat mempengaruhi akurasi. Pemilihan fitur adalah cara alternatif untuk mengurangi fitur yang tidak relevan, sehingga dimensi data berkurang dan kinerja klasifikasi menjadi tinggi. Fitur individual terbaik dapat dihasilkan dengan menggunakan beberapa metode, salah satunya adalah *Information Gain* (IG) [6]. Dalam penulisan jawaban dan umpan balik, Proses penilaian dapat dipersulit oleh kesalahan-kesalahan yang sering ditemukan. Kesalahan ini dapat mencakup kesalahan ejaan, aksen, dan tata bahasa dalam bahasa tertulis, yang membuat pemrosesan bahasa alami menjadi sulit [7]. Koreksi kata atipikal memiliki banyak algoritma dalam penerapannya. *Peter Norvig* adalah algoritma yang menggabungkan operasi penghapusan, penyisipan, substitusi, dan transposisi untuk kata-kata yang dikenali oleh kesalahan ketik dan mencarinya di kamus [8]. Sebelum klasifikasi, pembobotan fitur diperlukan untuk meningkatkan akurasi klasifikasi. Pembobotan yang umum digunakan adalah istilah *Term Frequency-Inverse Document Frequency* (TF-IDF) [9].

Berdasarkan penelitian sebelumnya tentang analisis sentimen berbasis perspektif pada wisata pantai di Malang Selatan, ditemukan bahwa metode untuk meningkatkan penggunaan kata-kata yang tidak biasa dengan tujuan mengoptimalkan jumlah fitur harus digunakan dalam langkah klasifikasi [10]. Selain itu, kajian tentang fungsi seleksi fitur dalam proses klasifikasi menunjukkan bahwa seleksi fitur berguna untuk mereduksi fitur agar proses klasifikasi dapat lebih efisien dan efektif [11]. Kemudian, penelitian yang berkaitan dengan normalisasi kata dalam *pre-processing* dan penggunaan seleksi fitur *Information Gain* menemukan bahwa dengan normalisasi meningkatkan hasil akurasi sebelumnya 94% pada pra-pemrosesan tanpa normalisasi kata dan sebesar 98% untuk *pre-processing* dengan kata-kata normalisasi dikombinasikan dengan *Information Gain* sebagai pemilihan fitur [12]. Selanjutnya, penelitian terkait *Naive Bayes* yang dikombinasikan dengan metode pemilihan fitur, yaitu *Information Gain* sebagai metode seleksi untuk memilih fitur yang efektif untuk setiap pengenal kelas, menunjukkan bahwa akurasi dan *f-measure* diperoleh dengan memilih fitur *Information Gain* pada klasifikasi yaitu 91,33% dan 89,18% [13]. Kemudian studi tentang penggunaan koreksi ejaan dengan metode *Peter Norvig* dan *N-Gram.* Hasil pengujian menunjukkan bahwa kombinasi kedua metode tersebut memberikan akurasi sebesar 73,684% dan tingkat kelulusan sebesar 37,037% untuk akurasi keseluruhan dari aplikasi ini adalah 69,09% [8]. Lalu kajian normalisasi kata oleh *Peter Norvig*, hasil penelitian ini menunjukkan normalisasi kata *Peter Norvig* menghasilkan *precision* 0,903, *recall* 0,944, *f-measure* 0,922 dan *accuracy* 0,903 [14].

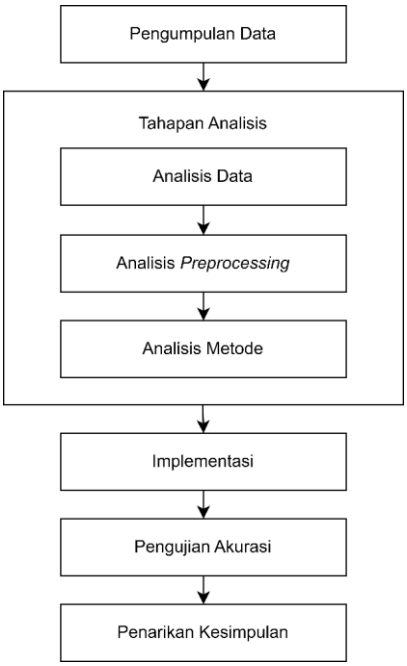
Penelitian ini berfokus pada latar belakang yang dijelaskan pada beberapa penelitian sebelumnya berdasarkan analisis sentimen yang dilakukan pada tingkat aspek dengan metode *Support Vector Machine* (SVM) dilanjutkan perbaikan pada kata yang tidak baku menggunakan algoritma *Peter Norvig* dengan Seleksi Fitur *Information Gain* untuk mengoptimalkan jumlah *feature*, kemudian dilakukan pembobotan data menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) yang akan digunakan pada analisis sentimen berbasis aspek terhadap *review* pariwisata pantai Malang Selatan.

## 2. METODOLOGI

Metodologi menjelaskan mengenai metode penelitian, dataset, analisis sentimen berbasis aspek, kata baku, arsitektur sistem, *case Folding, cleaning, tokenizing, normalization, stemming, convert negation, stopword removal*, korpus, n-gram, pembobotan tf-idf, *information gain*, algoritma svm, *kernel trick, k-fold cross validation, confusion matrix*, dan hasil pengujian.

### 2.1 Metode Penelitian

Pada penelitian ini terdapat lima tahapan alur penelitian, yaitu pengumpulan data, tahapan analisis, implementasi, pengujian akurasi, serta penarikan kesimpulan. Pada tahapan pengumpulan data, dilakukan pengumpulan data dari studi literatur yang berkaitan dengan penelitian dan pengumpulan dataset dari situs TripAdvisor dalam bentuk opini yang diberikan oleh pengunjung terhadap objek wisata pantai Malang Selatan menggunakan metode *Web Scraping*. Tahap berikutnya adalah tahapan analisis dilakukan analisis data, analisis *preprocessing*, dan analisis metode. Tahapan implementasi yaitu penerapan dari hasil analisis. Tahapan pengujian akurasi dilakukan pengukuran performansi akurasi dari SVM-IG dengan normalisasi kata *Peter Norvig*. Tahapan terakhir adalah penarikan kesimpulan yaitu berdasarkan hasil dari pengujian akurasi. Berikut merupakan blok diagram alur penelitian, dapat dilihat pada Gambar 1.




**Gambar 1.** Blok Diagram Alur Penelitian

### 2.2 Dataset

Data yang digunakan bersumber dari situs TripAdvisor menggunakan teknik Web Scraping untuk mendapatkan teks opini dengan menggunakan kata kunci pantai Malang Selatan pada tanggal 1 Januari 2013 sampai 1 Juni 2022 dengan jumlah total sebanyak 1020 ulasan. Terdapat lima aspek yang telah ditentukan yaitu umum, kebersihan, keramaian, akses jalan, kondisi ombak. Masing-masing aspek pada ulasan dinotasikan polaritas sentimennya, yaitu ”0” untuk polaritas netral, ”1” untuk polaritas yang memiliki sentimen positif, dan ”-1” untuk polaritas yang memiliki sentimen negatif.

### 2.3 Analisis Sentimen Berbasis Aspek

Analisis sentimen pada tingkat aspek juga dikenal sebagai *opinion mining* berbasis fitur, melakukan analisis sentimen secara lebih tepat dibandingkan dengan opini target. Mengidentifikasi polaritas sentimen (positif, negatif, atau netral) dari sebuah kalimat atau teks untuk sebuah objek yang merupakan bagian dari entitas tertentu adalah tujuan dari analisis sentimen berbasis aspek. Sistem sentimen berbasis aspek mengambil kumpulan teks tentang entitas tertentu sebagai masukan, mencoba memahami aspek atau karakteristik utama dari entitas tersebut, dan kemudian mengevaluasi sentimen rata-rata teks untuk setiap aspek [3]. Produk, layanan, topik, isu, orang, organisasi, dan peristiwa adalah ilustrasi dari sebuah entitas dalam analisis sentimen berbasis aspek, sedangkan aspek adalah elemen atau karakteristik dari entitas tersebut.

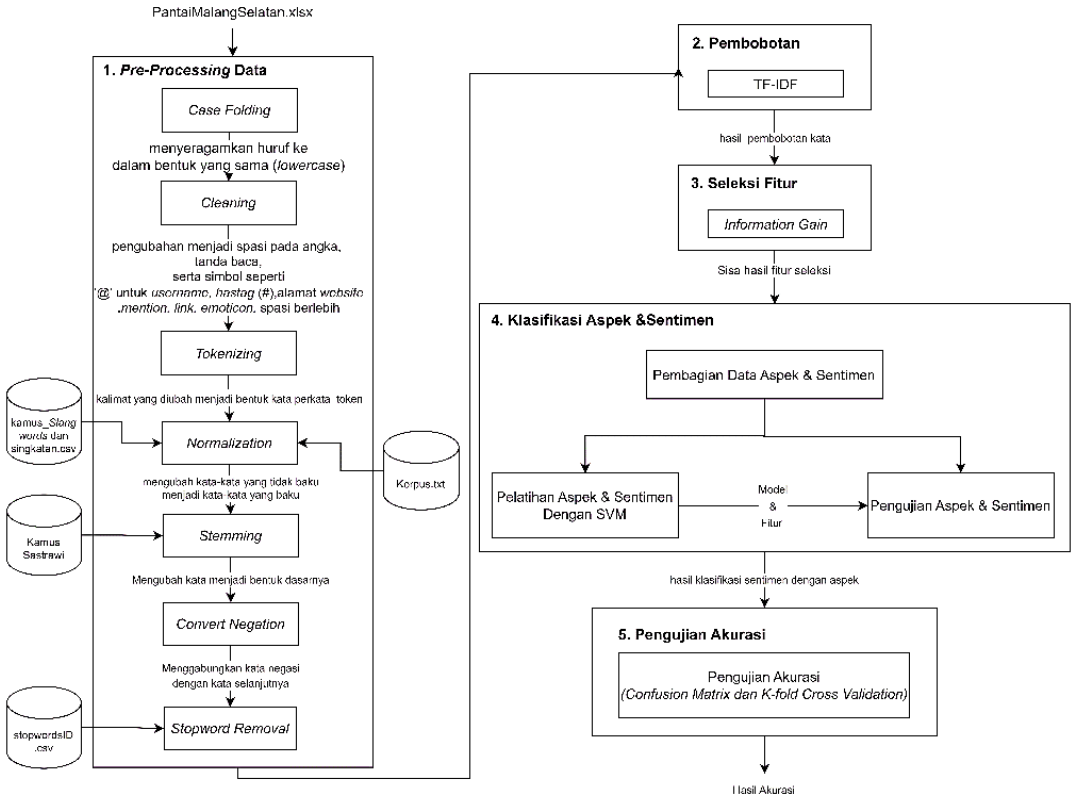
### 2.4 Kata Baku

Menurut kaidah penulisan kata bahasa Indonesia, penggunaan kata baku sangat penting dalam penulisan karya tulis ilmiah. Kata baku biasanya digunakan dalam kalimat formal atau dalam beberapa bahasa baku, termasuk bahasa lisan dan tulisan [15]. Suatu kata dapat disebut kata tidak baku apabila kata yang

digunakan tidak mengikuti kaidah bahasa Indonesia. Kata atipikal ini sering muncul dalam kehidupan kita sehari-hari. Beberapa contoh kesalahan ejaan bahasa Indonesia saat menulis kata tidak baku adalah: "tehknik" seharusnya "teknik", "sepak bola" seharusnya "sepakbola", "apotik" seharusnya "apotek".

### 2.5 Arsitektur Sistem

Sistem yang akan dibangun melewati beberapa tahapan yaitu memasukkan dataset, *pre-processing*, pembobotan dengan menggunakan TF-IDF, lalu seleksi fitur dengan *Information Gain*, proses klasifikasi menggunakan *Support Vector Machine*, proses keluaran hasil analisis sentimen berdasarkan aspek, dan pengujian akurasi.




**Gambar 2.** Blok Diagram Sistem

### 2.6 *Pre-Processing data*

*Pre-processing* data merupakan tahapan dimana data disiapkan dalam bentuk teks sebelum diproses untuk tahapan selanjutnya [10]. Tujuan dari proses ini adalah agar data lebih terstruktur dan siap untuk diolah. Proses pra-pemrosesan itu sendiri terdiri dari beberapa tahapan. Langkah-langkah persiapan yang harus dilakukan adalah sebagai berikut:

1. *Case Folding*, yaitu membuat semua huruf memiliki bentuk yang sama (*lowercase*).
2. *Cleaning*, yaitu pembersihan kata, tanda baca atau simbol yang tidak diperlukan pada proses hasil klasifikasi.
3. *Tokenizing*, yaitu pemotongan kata berdasarkan spasi.
4. *Normalization*, yaitu memperbaiki kesalahan kata atau singkatan ke dalam bentuk tertentu berdasarkan kamus. Pada tahapan ini menggunakan normalisasi kata dengan kamus *Slang Words* dan singkatan (SS) dilanjutkan normalisasi kata menggunakan *Peter Norvig* dengan korpus.
5. *Stemming*, yaitu Pengubahan kata imbuhan menjadi kata dasar.
6. *Convert Negation*, yaitu jika kata negasi ditemukan maka gabungkan dengan kata berikutnya.
7. *Stopword Removal*, yaitu menghilangkan kata yang terdapat di dalam daftar stopword.

### 2.7 Kamus *Slang Words* dan Singkatan (SS)

Setiap kata dalam kamus diperiksa untuk tahap normalisasi [14]. Kata yang benar akan digunakan sebagai gantinya jika kata yang sama ada dalam kamus. Kata yang telah melewati proses normalisasi kata dan dikenali di kamus SS menjadi kata baku.

### 2.8 *Peter Norvig*

*Peter Norvig* adalah sebuah algoritma yang dapat mengubah jarak kata atau kata-kata yang salah eja menjadi dua kata. Dibutuhkan beberapa pengeditan untuk mengubah satu kata menjadi kata lainnya. *Peter Norvig* dapat menggunakan semua operasi edit jarak jauh, termasuk menyisipkan, mengganti, mengubah urutan, dan menghapus, untuk membuat kata apa pun yang mungkin terdapat kesalahan ketik [8]. Setelah semua fitur mendapatkan kata kandidat, selanjutnya menghitung probabilitas setiap kata kandidat yang cocok dengan korpus. Rumus untuk perhitungan *Peter Norvig* dapat dilakukan pada persamaan (1) berikut:

$$correction(w) = \text{argmax}_{c \in candidates}\, P(c|w) \quad (1)$$

Dimana:

argmax : pemilihan kandidat yang memiliki probabilitas tertinggi.
c∈candidates : kandidat kata c dari kumpulan kandidat.
P(c) : probabilitas kandidat c muncul pada sebuah corpus dokumen.
P(w|c) : menunjukkan probabilitas bahwa kata w adalah teks yang dimaksud pada kandidat c.

Dengan Teorema Bayes, setara dengan persamaan (2) berikut:

$$correction(w) = \text{argmax}_{c \in candidates} \frac{P(c)P(c|w)}{P(w)} \quad (2)$$

Setelah menghitung probabilitas kata kandidat dalam korpus, kesalahan ejaan "Sya" paling mendekati kata "Saya" pada korpus. Proses perhitungan metode ini berlanjut hingga ditemukan kata yang paling mendekati.

### 2.9 Korpus

Korpus adalah kumpulan teks otentik, baik tertulis maupun transkrip, yang disimpan di komputer dan dianalisis menggunakan perangkat lunak yang dirancang untuk analisis korpus [16]. Korpus tersebut dapat digunakan untuk mengumpulkan informasi bahasa di Internet, khususnya media online, setelah itu datanya dapat dianalisis. Contoh kumpulan teks yang akan ditangani dalam korpus adalah buku, majalah internasional, majalah, surat kabar, artikel, dll.

### 2.10 *N-gram*

*N-gram* sering digunakan untuk berbagai masalah, seperti dalam pengenalan suara, koreksi kata terjemahan, pencarian string, prediksi kata, dan koreksi ejaan [8]. Token n kata dari sebuah kalimat digunakan untuk metode *N-gram*. *N-gram* sendiri dibedakan berdasarkan jumlah n, *unigrams* diwakili oleh n = 1, sedangkan *bigrams* diwakili oleh n = 2. Berikut ini adalah ilustrasi pemenggalan kalimat dengan metode *N-gram* dengan frasa "harus tetap waspada.":

1. *Unigrams* : harus, tetap, waspada
2. *Bigrams* : harus tetap, tetap waspada

Probabilitas sebuah *N-gram* dengan menggunakan *Maximum Likehood Estimation* (MLE) [17] adalah asumsi bahwa kemunculan suatu kata bergantung pada kemunculan kata sebelum dan sesudah kemunculannya dalam kalimat dengan menghitung jumlah kemunculan *N-gram* ke dalam korpus dan kemudian membaginya dengan nilai total sehingga nilainya antara 0 dan 1. Dapat dilihat pada persamaan (3) berikut:

$$P_1\left(W^i|W^{i-1}\right) = \frac{count\left(W^{i-1}W^i\right)}{\sum_m count\left(W^{i-1}W\right)} \quad (3)$$

Dengan menggunakan *Maximum Likehood Estimation* (MLE), maka dihasilkan rumus-rumus yang dapat dilihat pada persamaan (4), dan (5) berikut:

$$P_1\left(c_j^i\right) = \frac{count\left(c_j^i\right)}{\sum_r^{k_i} count\left(c_r^i\right)} \quad (4)$$

$$P_2\left(c_j^i|W^{i-1}\right) = \frac{count\left(W^{i-1}c_j^i\right)}{\sum_r^{k_i} count\left(W^{i-1}c_r^i\right)} \quad (5)$$

Dimana:

$P_1$ : Probabilitas untuk unigram

$P_2$ : Probabilitas untuk bigram

**2.11 Pembobotan TF-IDF**

Data yang telah diproses sebelumnya harus diubah menjadi bentuk numerik, oleh karena itu diperlukan pembobotan kata [9]. Sebuah prosedur yang dikenal sebagai TF-IDF (*Term Frequency - Inverse Document Frequency*) melibatkan pengubahan data tekstual menjadi data numerik untuk memberikan bobot pada setiap kata. Metode ini melibatkan pengulangan kejadian kata-kata dalam dokumen.

Metode yang digunakan untuk pembobotan kata TF-IDF dapat dilihat pada rumus persamaan (6) berikut:

$$TF * IDF(d,t) = TF(d,t) * log(\frac{N}{df_t}) \quad (6)$$

Keterangan :

$TF * IDF(d,t)$ : Pembobotan TF-IDF
$TF(d,t)$ : Frekuensi munculnya term t pada dokumen d
$N$ : Jumlah dari semua kumpulan dokumen
$df_t$ : Jumlah dari dokumen yang mengandung term t

**2.12 *Information Gain***

Information Gain adalah salah satu metode seleksi fitur yang sering digunakan oleh para peneliti untuk mengidentifikasi batas kepentingan atribut. Perbedaan antara nilai entropi sebelum dan sesudah pemisahan adalah nilai IG [18]. Nilai ini digunakan untuk menentukan atribut mana yang akan dibuang atau digunakan. Atribut yang memenuhi kriteria bobot digunakan kemudian dalam proses klasifikasi.

Rumus untuk mendapatkan perhitungan dari entropi dapat dilihat pada persamaan (7) berikut:

$$Entropy(S) = \sum_{i=1}^{m} p_i log_2(pi) \quad (7)$$

dimana:

$m$ : Jumlah kelas klasifikasi
$p_i$ : Jumlah proporsi sampel (peluang) untuk kelas i

Sedangkan rumus untuk *Information Gain* dari suatu atribut A, ditunjukkan pada persamaan (8) berikut:

$$Gain(S,A) = Entropy(S) - \sum_v \frac{|S_v|}{|S|} Entropy(S_v) \quad (8)$$

dimana:

$A$ : Variabel
$v$ : Nilai yang mungkin untuk variabel A
$|S_v|$ : Jumlah sampel untuk nilai v
$|S|$ : Jumlah sampel untuk nilai v

**2.13 *Support Vector Machine***

*Support Vector Machine* adalah teknik klasifikasi untuk menentukan Maximum Marginal Hyperplane (MMH) atau pembeda yang paling efektif untuk mencapai tingkat pemisahan kelas yang paling tinggi [19]. Suatu *edge* dapat didefinisikan sebagai jarak terpendek dari *hyperplane* ke salah satu sisi *edge* yang sama dengan jarak dari *hyperplane* ke sisi *edge* yang lain, asalkan kedua *edge* sejajar dengan *hyperplane*. Dapat dilihat pada persamaan (9) berikut:

$$y = ax + b \quad (9)$$

Persamaan garis lurus tersebut dituliskan pada persamaan (10) berikut:

$$y - ax - b = 0 \quad (10)$$

Dimana:

a : Kemiringan atau gradien (m)
b : Bilangan konstanta, a dan b merupakan bilangan real dan a tidak nol.

Dalam SVM, secara general sebuah *hyperplane* dinyatakan pada persamaan (11) berikut:

$$w \cdot x + b = 0 \quad (11)$$

Keterangan:

w : Nilai dari bidang normal
x : Data input
b : posisi bidang relatif terhadap pusat koordinat, dimana skalar b bisa bernilai negatif, nol, maupun positif.

Kemudian untuk memaksimalkan L terhadap $\alpha_i$dijelaskan pada persamaan (12) dengan memperhatikan batasan pada persamaan (13) berikut:

$\max_{\alpha} L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$ (12)

$\alpha_i \geq 0 (i = 1,2, \dots, l) \sum_{i,i=1}^{l} \alpha_i y_i = 0$ (13)

Setelah mendapatkan *Support Vector* dengan $\alpha_i$bernilai positif, maka selanjutnya menghitung w yang dapat dilihat pada persamaan (14) dan menghitung b yang dapat dilihat pada persamaan (15) berikut:

$w = \sum_{i=1}^{n} \alpha_i y_i x_i = 0$ (14)

$b = y_k - w^T x_k$ (15)

Berikutnya untuk menghitung kelas $\overline{x}$, dapat dilihat pada persamaan (16) berikut:

$F(\overline{x}) = W^T x + b$ (16)

### 2.14 *Kernel Trick*

Pada trik kernel, hanya perlu mengetahui fungsi kernel yang digunakan untuk mendefinisikan *support vector*. Jadi tidak perlu diketahui bentuk fungsi nonlinier ϕ. Secara garis besar, ada empat jenis fungsi inti yang dapat digunakan, yaitu:

1. Kernel Linier, dapat dilihat pada persamaan (17) berikut:
   $K(x, x_k) = X_k^T x$ (17)
2. Kernel Polynomial, dapat dilihat pada persamaan (18) berikut:
   $K(x, x_k) = X_k^T x + 1)^d$ (18)
3. Kernel Gaussian, dapat dilihat pada persamaan (19) berikut:
   $K(x, x_k) = exp\{-||x - x_k||_2^2/\sigma^2\}$ (19)
4. Kernel Sigmoid, dapat dilihat pada persamaan (20) berikut:
   $K(x, x_k) = \tanh[kx_k^T x + \theta]$ (20)

### 2.15 *K-fold Cross Validation*

*K-fold Cross Validation* adalah metode proses validasi untuk mengevaluasi kinerja pembelajaran mesin atau model pembelajaran mesin. Secara umum, validasi silang K-fold digunakan untuk memperkirakan akurasi karena biasnya yang relatif rendah [20].

### 2.16 *Confusion Matrix*

*Confusion matrix* adalah tabel yang berisi informasi tentang jumlah data uji yang benar atau jumlah data tes yang salah [21]. Dapat dilihat pada Tabel 1 berikut:

**Tabel 1.** *Confusion Matrix*

| **Confusion Matrix** | | **Kelas Prediksi** | | |
|---|---|---|---|---|
| | | **Positif** | **Negatif** | **Netral** |
| **Kelas Sebenarnya** | **Positif** | TPP | PFNeg | PFNet |
| | **Negatif** | NegFP | TNegNeg | NegFNet |
| | **Netral** | NetFP | NetFNeg | TNetNet |

Tingkat akurasi menjelaskan perbandingan jumlah dari prediksi benar. Dapat dilihat pada persamaan (21) berikut:

$$Accuracy = \frac{\text{TPP} + \text{TNegNeg} + \text{TNetNet}}{\text{TPP}+\text{PFNeg}+\text{PFNet}+\text{NegFP}+\text{TNegNeg}+\text{NegFNet}+\text{NetFP}+\text{NetFNeg}+\text{TNetNet}}$$ (21)

Keterangan:

1. **TPP (*True Possitive Positive*)**, merupakan jumlah dokumen yang dikategorikan dengan benar sebagai bagian dari kelas positif
2. **TNegNeg (*True Negative Negative*)**, merupakan jumlah dokumen yang dikategorikan dengan benar sebagai bagian dari kelas negatif.
3. **TNetNet (*True Netral Netral*)**, merupakan jumlah kelas yang dikategorikan dengan benar sebagai bagian dari kelas netral.
4. **PFNeg (*Positive False Negatif*)**, merupakan jumlah dokumen kelas positif yang salah diklasifikasikan sebagai dokumen kelas negatif.
5. **NegFP (*Negatif False Positive*)**, merupakan jumlah dokumen negatif yang salah diklasifikasikan sebagai dokumen kelas positif.
6. **PFNet (*Positive False Netral*)**, merupakan jumlah dokumen positif yang salah diklasifiasikan sebagai dokumen kelas netral.
7. **NetFP (*Netral False Positive*)**, merupakan jumlah dokumen netral yang salah diklasifikasikan sebagai dokumen kelas positif..

8. **NetFNeg (*Netral False Negatif*)**, merupakan jumlah dokumen kelas netral yang salah diklasifikasikan sebagai dokumen kelas negatif.
9. **NegFNet (*Negatif False Netral*)**, merupakan jumlah dokumen kelas negatif yang salah diklasifikasikan sebagai dokumen kelas netral.

## 3. HASIL DAN PEMBAHASAN

Dataset yang digunakan untuk ulasan pariwisata pantai Malang Selatan pada pengujian ini yaitu sebanyak 1020 ulasan. Kemudian untuk polaritas dari tiap aspek dapat dilihat pada Tabel 2 berikut:

**Tabel 2.** Polaritas Tiap Aspek

| | | Polaritas | | | Jumlah |
|---|---|---|---|---|---|
| | | Positif | Negatif | Netral | |
| Aspek | Umum | 930 | 58 | 32 | 1.020 |
| | Kebersihan | 164 | 114 | 742 | 1.020 |
| | Keramaian | 90 | 40 | 890 | 1.020 |
| | Akses Jalan | 104 | 214 | 702 | 1.020 |
| | Kondisi Ombak | 116 | 132 | 772 | 1.020 |

Dataset akan dibagi menjadi data latih dan data uji dengan menggunakan representasi *k-fold cross validation* sebesar 10 fold, data yang digunakan sebagai data masukan pada proses SVM *training* sebanyak 816 data, kemudian data yang digunakan sebagai *testing* sebanyak 204 data.

### 3.1. Hasil Pengujian *Support Vector Machine* Tanpa *Information Gain* dan *Peter Norvig*

*Klasifikasi Support Vector Machine* tanpa *information gain* dan *Peter Norvig* hasil rata-rata akurasinya dapat dilihat pada Tabel 3 berikut:

**Tabel 3.** Hasil Rata-Rata Akurasi *Support Vector Machine* Tanpa *Information Gain* dan *Peter Norvig*

| Aspek | Akurasi |
|---|---|
| Umum | 0,90 |
| Kebersihan | 0,79 |
| Keramaian | 0,90 |
| Akses Jalan | 0,74 |
| Kondisi Ombak | 0,77 |
| **Hasil Rata-Rata Akurasi** | 0,82 |

Klasifikasi *Support Vector Machine* tanpa *Information Gain* dan *Peter Norvig* memperoleh rata-rata akurasi sebesar 0,82. Dengan akurasi terbesar yaitu 0,90 yang didapat dari aspek umum dan aspek keramaian.

### 3.2. Hasil Pengujian *Support Vector Machine* Tanpa *Information Gain*

*Klasifikasi Support Vector Machine* tanpa *information gain* hasil rata-rata akurasinya dapat dilihat pada Tabel 4 berikut:

**Tabel 4.** Hasil Rata-Rata Akurasi *Support Vector Machine* Tanpa *Information Gain*

| Aspek | Akurasi |
|---|---|
| Umum | 0,89 |
| Kebersihan | 0,76 |
| Keramaian | 0,89 |
| Akses Jalan | 0,77 |
| Kondisi Ombak | 0,77 |
| **Hasil Rata-Rata Akurasi** | 0,82 |

Klasifikasi *Support Vector Machine* tanpa *Information Gain* memperoleh rata-rata akurasi 0,82. Dengan akurasi terbesar yaitu 0,89 yang didapat dari aspek umum dan aspek keramaian.

### 3.3. Hasil Pengujian *Support Vector Machine* Tanpa *Peter Norvig*

*Klasifikasi Support Vector Machine* tanpa *peter norvig* hasil rata-rata akurasinya dapat dilihat pada Tabel 5 berikut:

**Tabel 5.** Hasil Rata-Rata Akurasi *Support Vector Machine* Tanpa *Peter Norvig*

| Aspek | Akurasi |
|---|---|
| Umum | 0,91 |
| Kebersihan | 0,82 |
| Keramaian | 0,87 |
| Akses Jalan | 0,79 |
| Kondisi Ombak | 0,74 |
| **Hasil Rata-Rata Akurasi** | 0,83 |

Klasifikasi *Support Vector Machine* tanpa *Peter Norvig* memperoleh rata-rata akurasi 0,83. Dengan akurasi terbesar yaitu 0,91 yang didapat dari aspek umum.

### 3.4. Hasil Pengujian *Support Vector Machine* Dengan *Information Gain* dan *Peter Norvig*

Klasifikasi *Support Vector Machine* Dengan *Information Gain* dan *Peter Norvig* rata-rata akurasinya dapat dilihat pada Tabel 6 berikut:

**Tabel 6.** Hasil Rata-Rata Akurasi *Support Vector Machine* Dengan *Information Gain* dan *Peter Norvig*

| Aspek | Akurasi |
|---|---|
| Umum | 0,92 |
| Kebersihan | 0,82 |
| Keramaian | 0,85 |
| Akses Jalan | 0,75 |
| Kondisi Ombak | 0,73 |
| **Hasil Rata-Rata Akurasi** | 0,81 |

Klasifikasi *Support Vector Machine* dengan *Information Gain* dan *Peter Norvig* memperoleh rata-rata akurasinya 0,81. Dengan akurasi terbesar yaitu 0,92 yang didapat dari aspek umum.

### 3.5. Pembahasan Pengujian

Dari hasil pengujian terhadap data uji yang telah dilakukan, hasil akurasi yang didapatkan adalah sebagai berikut pada Tabel 7 berikut:

**Tabel 7.** Pembahasan Pengujian

| No | Pengujian | Akurasi Aspek | | | | | Rata-Rata Akurasi |
|---|---|---|---|---|---|---|---|
| | | Umum | Kebersihan | Keramaian | Akses Jalan | Kondisi Ombak | |
| 1 | SVM | 0,90 | 0,79 | 0,90 | 0,74 | 0,77 | 0,82 |
| 2 | SVM + *Peter Norvig* | 0,89 | 0,76 | 0,89 | 0,77 | 0,77 | 0,82 |
| 3 | SVM + *Information Gain* | 0,91 | 0,82 | 0,87 | 0,79 | 0,74 | 0,83 |
| 4 | SVM + *Information Gain* + *Peter Norvig* | 0,92 | 0,82 | 0,85 | 0,75 | 0,73 | 0,81 |

## 4. PENUTUP

Berdasarkan hasil keseluruhan proses yang sudah dilakukan dalam penelitian pengaruh *Information Gain* dengan perbaikan kata tidak baku menggunakan kamus *slang words* dan singkatan juga *spelling correction* menggunakan metode *Peter Norvig* kata pada analisis sentimen berbasis aspek didapatkan kesimpulan bahwa analisis sentimen berbasis aspek dengan *Information gain* tanpa normalisasi kata *Peter Norvig* menghasilkan akurasi yang lebih baik yaitu sebesar 83% dibandingkan dengan algoritma *Support Vector Machine* dengan *Information Gain* dan normalisasi kata *Peter Norvig* sebesar 81%. Penggunaan normalisasi kata *Peter Norvig* tanpa *Information Gain* memperoleh akurasi rata-rata 82% dan klasifikasi *Support Vector Machine* tanpa normalisasi kata *Peter Norvig* dan *Information Gain* menghasilkan rata-rata akurasi sebesar 82%. Penurunan hasil akurasi tersebut disebabkan kesalahan dalam pengubahan kata karena kata yang dapat diubah tersebut hanya dapat mengoreksi 1 huruf yang salah.

Dari penelitian yang telah dilakukan, dapat dilakukan pengembangan lebih lanjut agar menjadi lebih baik kedepannya. Program ini sangat bergantung pada kualitas kamus daftar kalimat, sehingga kamus kalimat harus mencakup banyak kalimat yang umum dan menambahkan metode normalisasi kata yang dapat memperbaiki perubahan jarak huruf yang lebih dari 2 untuk mendapatkan hasil evaluasi yang lebih baik. Mengenai saran yang dapat diberikan agar sistem berfungsi lebih baik pada pengembangan selanjutnya, diantaranya pada penelitian selanjutnya diharapkan untuk menambah jumlah data yang digunakan karena akan mempengaruhi proses identifikasi untuk mendapatkan hasil yang optimal.

## DAFTAR PUSTAKA

[1] W. Parasati, F. A. Bachtiar, dan N. Y. Setiawan, "Analisis Sentimen Berbasis Aspek pada Ulasan Pelanggan Restoran Bakso President Malang dengan Metode Naïve Bayes Classifier", Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol. 4 No. 4, pp 1090-1099, 2020.

[2] Y. Susandi, A. Herdiani, dan I. L. Sardi "Analisis Sentimen Opini Masyarakat Terhadap Pasangan Calon Presiden dan Wakil Presiden pada Media Sosial Twitter Menggunakan Ontology Supported Polarity Mining", E-Proceeding of Engineering Vol. 6 No. 2, pp 8670-8681, 2019.

[3] Nuryani dan D. Mahayana "Analisis Sentimen Berbasis Aspek dengan Deep Learning Ditinjau dari Sudut Pandang Filsafat Ilmu", Jurnal Masyarakat Informatika Unjani Vol. 4 No. 2, pp 70-85, 2021.

[4] R. Sari dan R. Y. Hayuningtyas, "Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Pada Wisata TMII Berbasis Website", Indonesian Journal on Software Engineering Vol. 5 No. 2, pp 51-60, 2019.

[5] A. S. Ritonga dan E. S. Purwaningsih, "Penerapan Metode Support Vector Machine (SVM) Dalam Klasifikasi Kualitas Pengelasan SMAW (Shield Metal ARC Welding", Journal Ilmiah Edutic Vol. 5 No. 1, pp 17-25, 2018.

[6] M. H. A. Nurjaman, M. S. Mubarok, dan Adiwijaya "Analisis Sentimen Pada Ulasan Buku Berbahasa Inggris Menggunakan *Information Gain* dan *Support Vector Machine*", *E- Proceeding of Engineering* Vol. 4 No. 3, pp 4900-4906, 2017.

[7] Y. M. Febrianti, "Analisis Sentimen Pada Ulasan Lazada Berbahasa Indonesia Menggunakan K-Nearest Neighbor (K-NN) Dengan Perbaikan Kata Menggunakan Jaro Winkler Distance" *Other thesis*, Universitas Brawijaya, 2018.

[8] R. Martin, D. S. Naga, dan V. C. Mawardi, "Penggunaan Spelling Correction Dengan Metode Peter Norvig dan N-Gram ", Jurnal Ilmu Komputer dan Sistem Informasi Vol. 9 No. 1, pp 175-180, 2021.

[9] A. E. Irsad, Y. A. Sari, dan M. A. Fauzi, "Seleksi Fitur Information Gain untuk Klasifikasi Informasi Tempat Tinggal di Kota Malang Berdasarkan Tweet Menggunakan Metode Naïve Bayes dan Pembobotan TF-IDF-CF", Jurnal Pengembangan Teknologi dan Ilmu Komputer Vol. 3 No. 5, pp 4907-4913, 2019.

[10] Y. T. Pratama, F. A. Bachtiar, dan N. Y. Setiawan, "Analisis Sentimen Opini Pelanggan Terhadap Aspek Pariwisata Pantai Malang Selatan Menggunakan TF-IDF dan Support Vector Machine", Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol. 2 No. 12, pp 6244-6252, 2018.

[11] S. Fanissa, M. A. Fauzi, dan S. Adinugroho, "Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking" Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol. 2 No. 8, pp 2766-2770, 2018.

[12] Ivan, Y. A. Sari, dan P. P. Adikara, "Klasifikasi Hate Speech Berbahasa Indonesia di Twitter Menggunakan Naive Bayes dan Seleksi Fitur Information Gain dengan Normalisasi Kata" Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol. 3 No. 5, pp 4914-4922, 2019.

[13] A. E. Sari, S. Widowati, dan K. M. Lhaksmana, "Klasifikasi Ulasan Pengguna Aplikasi Mandiri Online di Google Play Store dengan Menggunakan Metode Information Gain dan Naive Bayes Classifier" e-Proceeding of Engineering Vol. 6 No. 2, pp 9143-9157, 2019.
[14] T. M. Iryana, Indriati, dan P. P. Adikara, "Analisis Sentimen Masyarakat Terhadap Mass Rapid Transit Jakarta Menggunakan Metode Naïve Bayes Dengan Normalisasi Kata", Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol. 5 No 6, pp 2753-2760, 2021.
[15] T. M. Fahrudin, I. Sa'diyah, dan Latipah, "Aplikasi Pendeteksi Kesalahan Ejaan Bahasa Indonesia pada Karya Ilmiah Bidang Ilmu Komputer Menggunakan KEBI 1.0 Checker" Seminar Nasional Informatika Bela Negara Vol. 2, pp 64-72, 2021.
[16] M. R. Alimuddin, Gusnawaty, A. A. Salim, "Makna Stance Expressions dalam Teks Jurnalistik Media Berita Detik dan Kompas pada Topik Perubahan Iklim di Indonesia: Analisis Linguistik Korpus" Jurnal Sinestesia Vol. 12, No. 1, pp. 109-123, 2022
[17] S. F. Chen and J. Goodman, "Smoothing Techniques for Language Modeling" Center for Research in Computing Technology, 1998.
[18] Jihad, N. I. Widiastuti, dan K.E. Dewi, "Support Vector Machine Untuk Ekstraksi Dokumen Karya Ilmiah" KOMPUTA:Jurnal Ilmiah Komputer dan Informatika, Vol. 10, No. 2, pp 87-94, 2021.
[19] Suyanto, "Machine Learning Tingkat Dasar dan Lanjut" Informatika Bandung, 2018.
[20] J. Han, M. Kamber, and J. Pei "Data Mining Concepts and Techniques Third Edition" Elsevier Inc, 2012.
[21] D. Normawati dan S. A. Prayogi, "Implementasi Naive Bayes Classifier dan Confusion Matrix Pada Analisis Sentimen BerbasisTeks Pada Twitter", Jurnal Sains Komputer dan Informatika , Vol. 5, No. 2 , PP. 697-711, 2021.