Analisis Sentimen pada Produk Kecantikan dari Ulasan *Female Daily* Menggunakan Information Gain dan *SVM Classifier*

Nadifa Fadila Putri[1], Said Al Faraby[2], Mahendra Dwifebri[3]

[1,2,3] Universitas Telkom, Bandung
nadifadilaputri@students.telkomuniversity.ac.id[1], saidalfaraby@telkomuniversity.ac.id[2], mahendradp@telkomuniversity.ac.id[3]

Abstrak

Dampak dari perkembangan teknologi salah satunya yaitu pembelian produk secara online semakin disukai oleh masyarakat. Ulasan produk dapat membantu konsumen dalam mengetahui kualitas pada produk tersebut. Analisis sentimen adalah sebuah proses menemukan opini untuk menentukan pendapat atau sikap seseorang atas produk, layanan, atau organisasi lalu mengidenfikasi ke dalam sentimen yang telah diungkapkan dan kemudian mengklasifikasikannya. Oleh karena itu, pada penelitian ini akan melakukan penelitian terhadap klasifikasi sentiment untuk mencari performansi terbaik ulasan produk kecantikan dengan menggunakan seleksi fitur *Information Gain menggunakan* dataset yang bersumber dari situs *Female Daily,* dengan dilakukan beberapa proses yaitu melakukan *preprocessing,* pada ekstraksi fitur menggunakan *TF-IDF* untuk mengetahui jumlah bobot dari setiap kata, dan *feature selection* menggunakan metode *Information gain* untuk mereduksi dimensi pada sebuah dataset dengan menghapus atau mengurangi fitur-fitur yang dianggap tidak diperlukan. Hasil dari penelitian ini mendapat akurasi sebesar 85.89%, dan klasifikasi menggunakan algoritma *Support Vector Mechine* (SVM) mendapat akurasi sebesar 85.98% dengan menggunakan kernel sigmoid.

Kata kunci : Analisis Sentimen, Ulasan Produk, *Term Frequency Inverse Document Frequency (TFIDF),* Information Gain, Support Vector Machine

Abstract

The impact of technological developments is one of them is the purchase of products online increasingly preferred by the public. Product reviews can help consumers in knowing the quality of the product. Sentiment analysis is the process of finding an opinion to determine a person's opinion or attitude over a product, service, or organization and then identifying it into the sentiments that have been expressed and then classifying them. Therefore, this study will conduct research on sentiment classification to find the best performance of beauty product reviews by using *information gain* feature selection using datasets sourced from the Female Daily *site,* by doing several processes that *preprocessing,*on the extraction of features using *TF-IDF* to find out the amount of weight of each word, and *feature selection* using the *Information gain* method to reduce the dimensions of a dataset by removing or reducing features that are considered unnecessary. The results of this study got an accuracy of 85.89%,and classification using the Support Vector *Mechine* (SVM) algorithm got an accuracy of 85.98% withsigmoid kernels.

Keywords: ***Sentiment Analysis, Product Reviews, Term Frequency Inverse Document Frequency (TFIDF),, Information Gain, Support Vector Machine***

1. Pendahuluan

Latar Belakang

Pada saat ini perkembangan teknologi dan penerapan Internet telah membawa revolusi kepada masyarakat[1]. Pembelian produk secara online semakin disukai dan telah menjadi trending untuk membeli berbagai macam produk salah satunya yaitu produk kecantikan[2]. Ulasan produk kecantikan dapat membantu konsumen dalam mengetahui kualitas pada produk kecantikan tersebut layak atau tidak untuk digunakan. Namun, tidak semua produk kecantikan memiliki kualitas yang baik sesuai kebutuhan konsumen dan hal ini yang harus diperhatikan oleh para konsumen[3].

Dengan adanya analisis sentiment dibutuhkan tahap klasifikasi untuk mengklasifikasikan sentiment menjadi sentiment positif, negatif atau sentiment netral[2]. Telah dibuktikan pada penelitian sebelumnya yaitu pada penelitian [4] bahwa klasifikasi menggunakan SVM menghasilkan nilai rata-rata akurasi sebesar 85% menunjukan bahwa hasil dari metode SVM optimal. Oleh karena itu, pada penelitian ini penulis akan melakukan penelitian terhadap klasifikasi sentiment untuk mencari performansi terbaik ulasan produk kecantikan dengan menggunakan seleksi fitur *Information Gain*.

Pada penelitian [5] Algoritma *Information Gain* menghasilkan nilai rata-rata akurasi tertinggi dengan persentase sebesar 82,19%. Dapat disimpulkan bahwa *Information Gain* merupakan metode untuk seleksi fitur terbaik dalam analisis sentiment, dan keuntungan *Information* Gain menjadi metode yang dipilih untuk dijadikan sebagai algoritma pemilihan fitur. Maka, pada penelitian ini penulis menerapkan algoritma SVM dan algoritma *Information Gain* sebagai seleksi fitur untuk mendapatkan hasil akurasi yang tinggi. Berdasarkan kondisi yang telah dijelaskan, diperlukan analisis sentiment pada ulasan produk kecantikan yang akan menggunakan metode *Support Vector Machine* (SVM), metode ini dipilih karena dapat menghasilkan hasil yang optimal dalam klasifikasi[6].

Topik dan Batasannya

Pada penelitian ini penulis akan membangun model untuk analisis sentiment terhadap ulasan produk kecantikan. Penelitian ini berfokus terhadap proses *preprocessing*, ekstraksi fitur, seleksi fitur, dan klasifikasi. Pada proses *preprocessing*, penulis akan membandingkan pengaruh pada proses *preprocessing.* Pada proses ekstraksi fitur penulis menggunakan TF-IDF, dan seleksi fitur menggunakan *Infromation Gain.* Lalu pada proses klasifikasi menggunakan metode SVM. Penelitian ini menggunakan dataset yang diambil dari situs *femaledaily.com* dengan jumlah 3690 data menggunakan ulasan berbahasa Indonesia dan Bahasa inggris, dan terdapat beberapa aspek yaitu *price, packaging, product* dan aroma. Dataset ini menggunakan label dengan kelas sentimen positif, negatif, dan netral.

Tujuan

Tujuan dilakukannya penelitian ini yaitu untuk melakukan proses analisis sentimen dari suatu ulasan produk dengan menggunakan metode SVM dan membandingkan kernel linear, rbf, dan sigmoid, serta menganalisis permoformansi seleksi fitur dengan menggunakan *information gain* terhadap algoritma SVM dari suatu ulasan produk. Penelitian ini akan membandingkan pengaruh pada proses *preprocessing* terhadap proses analisis sentimen dari suatu ulasan produk kecantikan.

Organisasi Tulisan

Bagian selanjutnya pada penelitian ini adalah bagian 2 yang membahas mengenai studi terkait dengan penelitian yang dilakukan, bagian 3 membahas rancangan sistem yang dibangun, bagian 4 membahas evaluasi dari hasil pengujian, dan bagian 5 membahas kesimpulan dari penelitian ini dan saran untuk penelitian selanjutnya.

2. Studi Terkait

2.1 Penelitian Terkait

Dalam penelitian ini, penulis mengambil beberapa referensi mengenai penelitian sebelumnya terkait analisis sentimen pada produk. Adapun beberapa penelitian yang terkait yaitu:

Penelitian yang dilakukan oleh Mr. Nilesh M.Shelke, Dr. Shriniwas Deshpande [7] Pada tahun 2016, dengan judul "*Identification of Scope of Valence Shifter for Sentiment Analysis of Product Reviews"* yang bertujuan untuk memahami pembalikan polaritas seperti "baik" dan "tidak semua baik" serta karena asosiasi yang tepat dari intensiers dan pereda, telah membantu untuk menentukan polaritas dan intensitas fitur sentimen yang akurat. Pada penelitian ini dilakukan penanganan pemindah kontekstual seperti negasi dan pemindah valensi atau peredam atau intensifiers masih dalam tahap awal. Hasil yang didapatkan pada penelitian ini metode *naïve bayes* mendapatkan hadil sebesar 85%, menggunakan SVM sebesar 85,2%, menggunakan *genetic algoritma* sebesar 85,3%, dan terakhir menggunakan metode *proposed system* mendapatkan hasil sebesar 87%.

Pada penelitian [3] yang dilakukan oleh Dinar Ajeng Kristiyanti, pada tahun 2015 dengan judul "Analisis Sentimen Review Produk Kosmetik Melalui Komparasi *Feature* Selection" yang bertujuan untuk mengatasi masalah dengan cara otomatis mengkelompokan review pengguna menjadi opini positif atau negatif. Pada penelitian ini dilakukan pembandingan metode SVM & *Particle Swarm Optimization* dengan Svm & *Genetic Algoritma* untuk mengklasifikasi teks pada review produk kosmetik dalam meningkatkan akurasi Analisa sentimen. Hasil yang didapatkan pada penelitian ini metode SVM & *Particle Swarm Optimization* mendapatkan hasil lebih besar yaitu 97%, sedangkan jika menggunakan metode Svm & *Genetic Algoritma* mendapat hasil yaitu94%. Dapat disimpulkan bahwa metode SVM & *Particle Swarm Optimization* memberikan hasil akurasi yang lebih tinggi.

Penelitian yang dilakukan oleh Sari Widya Sihwi, Insan Prasetya jati dan Rini Anggainingsih [5] pada tahun 2018, dengan judul *"Twitter Sentiment Analysis of Movie Reviews Using Information Gain and Naïve Bayes Classifier"* yang bertujuan untuk mengetahui apakah tweet tersebut opini positf, opini negative, atau opini netral. Pada penelitian ini menggunakan metode algoritma naïve bayes classifier. Dataset yang digunakan dalam penelitian ini yaitu dari tweet 12 judul film popular. Hasil yang didapatkan dengan metode naïve bayes classifier dan information gain dalam penenelitian ini yaitu mendapatkan akurasi 82.19% dengan 0,006 sebagai ambang gain optional.

Penelitian yang dilakukan oleh Novelty Octaviani Faomasi Daeli [8] pada tahun 2019 dengan judul *"Sentiment analysis on movie reviews using Information gain and K-nearest neighbor*" bahwa pada penelitian ini, Polaritas v2.0 dari dataset review film Cornell akan digunakan untuk menguji KNN dengan pemilihan fitur Information gain untuk mencapai kinerja yang baik. Tujuan dari penelitian ini adalah menemukan K yang optimal untuk KNN berdasarkan ambang IG, dan menemukan ambang Information Gain terbaik.

2.2 Analisis Sentimen

Analisis Sentimen adalah proses untuk menentukan pendapat atau sikap seseorang atas produk, layanan, atau organisasi [2] Target dari analisis sentimen adalah menemukan opini, lalu mengidentifikasi sentimen yang telah diungkapkan dan kemudian mengklasifikasikannya.

Analisis sentimen mengklasifikasikan sebuah kalimat menjadi kelas positif dan kelas negative. Analisis sentimen dibagi menjadi 3 tingkat yaitu, tingkat dokumen, tingkat kalimat dan tingkat aspek. Analisis sentimen tingkat dokumen bertujuan untuk mengklasifikasikan dokumen opini seperti mengungkapkan sentimen positif atau negative, yang menganggap seluruh dokumen sebgai unit dasar informasi. Analisis sentimen tinkat kalimat bertujuan untuk mengklasifikasikan sentimen yang diekspresikan setiap kalimat, yang dilakukan pertama kali yaitu mengidentifikasi apakah kalimat tersebut subjektif atau objektif. Analisis sentimen tingkat aspek bertujuan untuk mengklasifikasikan sentimen terkait dengan aspek spesifik dari entitas, yang dilakukan pertama kali yaitu mengidentifikasi entitas dan aspeknya [9].

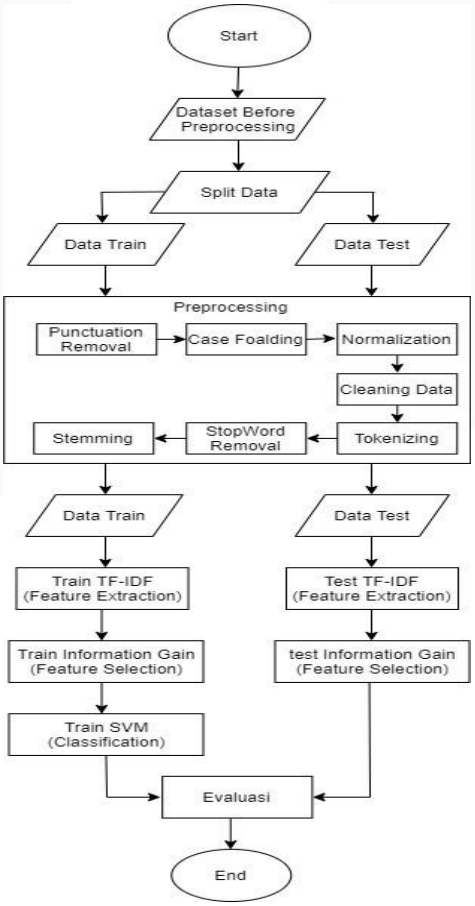
2.3 Ulasan Produk

Ulasan produk adalah sebuah tulisan yang menerangkan tentang keadaan suatu produk yang telah dibeli oleh pembeli. Ulasan produk biasanya ditulis pada kolom yang sudah disiapkan oleh pengembang aplikasi E-commerce, youtube, blog atau sosial media lainnya, berisi tulisan yang mencerminkan kualitas nyata dari suatu produk tersebut[10].

3. Sistem yang Dibangun

3.1 Skema Umum

Pada penelitian ini akan menggunakan *feature selection* yaitu *information gain*, dan akan dipasangkan dengan menggunakan metode *Support Vector Machine* (SVM). Data set akan dibagi menjadi 2 bagian yaitu 80% digunakan menjadi data latih, dan 20% akan digunakan menjadi data uji. Berikut merupakan skema umum yang akan dibangun:



Dari flowchart diatas dapat dilihat bahwa, tahap pertama yang akan dilakukan adalah membaca dataset *review product* yang akan digunakan pada penelitian. Tahap kedua adalah *split dataset* untuk membagi data menjadi data latih dan data uji. Tahap ketiga adalah *preprocessing* dilakukannya proses pembersihan dan penyiapan teks untuk tahapan selanjutnya. Tahap keempat adalah *Feature Extraction.* Tahap kelima adalah *Feature Selection* untuk mereduksi dimensi pada sebuah data dengan menghapus atau mengurangi fitur-fitur yang dianggap tidak diperlukan. Tahap keenam adalah *Classification* untuk melatih data uji. Tahapan terakhir adalah *Evaluation Matrix* untuk melakukan perhitungan tingkat akurasi dari data prediksi yang telah dilatih pada klasifikassi.

3.2 Dataset

Dataset yang digunakan dalam penelitian ini yaitu diambil dari kumpulan ulasan produk kecantikan dari situs *Female Daily*. Ulasan produk kecantikan pada dataset ini berupa produk kecantikan seperti *skincare*, menggunakan ulasan berbahasa Indonesia dan Bahasa inggris, namun dataset ini lebih dominan menggunakan bahasa Indonesia. Total data dari dataset tersebut berjumlah 3690 data dan mempunyai 6 atribut, yaitu *review_id, review_text, price, packaging, product,* dan aroma, dan pada dataset ini telah dilabeli secara manual oleh penelitian sebelumnya [11] kedalam 4 aspek yaitu aspek *price, packaging, product,* dan aroma. Label 1 digunakan untuk aspek label positif, label -1 digunakan untuk label negatif, dan label 0 digunakan untuk label netral. Berikut merupakan contoh ulasan produk kecantikan dan distribusi aspek dari dataset penelitian ini:

Tabel 1. Dataset yang digunakan

Teks Ulasan Produk	Price	Packaging	Product	Aroma
pertama kali pake ini, di kulit enak banget..udah gitu dia ga se-oily sunblock lainnya...cocok buat dipake sebelum bedak buat sehari2..harganya emg rada mahal tp buat gue awet sampe 6 bulanan jd worth it bgt..	-1	0	1	0

Pada data penelitian ini setiap aspek data diberikan label, Adapun pendistribusian aspek pada masing-masing label :

Tabel 2. Jumlah masing-masing label di setiap kelas

Aspek	Sentimen			Total Data Setiap Aspek
	-1	0	1	
Price	716 (18%)	2187 (55.22%)	1057 (26.96%)	3960
Packaging	189 (4.77%)	3324 (83.93%)	447 (11.28%)	3960
Product	688 (17.37%)	660 (16.66%)	2612 (65.95%)	3960
Aroma	218 (5.5%)	3073 (77.6%)	669 (16.89%)	3960

3.3 Split Data

Pada proses split data dilakukan pembagian data, data dibagi menjadi 2 bagian yaitu 80% digunakan menjadi data *training* , dan 20% akan digunakan menjadi data uji. Pada proses ini split data menggunakan *random state,* tujuan *random state* ini yaitu untuk membuat nilai konsisten saat sistem dijalankan. Adapun jumlah data train dan data test dari setiap label dan setiap kelas sebagai berikut :

Tabel 3. **Jumlah masing-masing label di setiap data**

Aspek	Data *Train*	Total	Data *Test*	Total

	-1	0	1		-1	0	1	
Price	564 (17.80%)	1768 (55.80%)	836 (26.38%)	3.168	152 (19.19%)	419 (52.90%)	221 (27.90%)	792
Packaging	148 (4.67%)	2652 (83.71%)	368 (11.61%)	3.168	41 (5.17%)	672 (84.84%)	79 (9.97%)	792
Product	553 (17.45%)	539 (17.01%)	2076 (65.53%)	3.168	135 (17.04)	121 (15.27%)	536 (67.67%)	792
Aroma	178 (5.61%)	2458 (77.58%)	532 (16.79%)	3.168	40 (5.05%)	615 (77.65%)	137 (17.29%)	792

3.4 Preprocessing

Dataset yang telah dipersiapkan sebelumnya harus melalui tahap preprocessing terlebih dahulu. Proses preprocessing dilakukan untuk mendapatkan apakah data tersebut berkualitas atau tidak. Adapun beberapa proses yang dilakukan yaitu:

1. Punctuation Removal
   Pada tahap ini, dilakukan pembersihan data dengan menghapus tanda baca, angka atau karakter khusus selain kata pada dataset[12]
2. Case Folding
   Pada tahap ini, seluruh huruf pada data akan diubah menjadi huruf kecil (lower case)[13].
3. *Normalization*
   Pada tahap ini, terjadi proses pengubahan kata yang tidak standar menjadi kata standar dengan membuat kamus data secara manual.
4. *Data Cleaning*
   Pada tahap ini, terjadi proses penghapusan tanda atau simbol, dan juga menghapus angka.
5. Tokenizing
   Pada tahap ini, terjadi proses pemisahan teks menjadi potongan-potongan yang disebut sebagai token untuk kemudian di analisa [13].
6. Stopword Removal
   Pada tahap ini, terjadi penghapusan kata-kata yang dianggap tidak dapat memberikan pengaruh dalam menentukan suatu kategori sentimen seperti : 'yang', 'untuk', 'pada', 'ke', 'para', 'namun', 'menurut', 'antara', 'dia', 'dua', 'ia', 'seperti', 'jika', 'jika', 'sehingga', 'kembali', 'dan', 'tidak', 'ini', 'karena', 'kepada', 'oleh', 'saat', 'harus', 'sementara', 'setelah', 'belum', 'kami', dan lain-lain [12].
7. Stemming
   Pada tahap ini, terjadi proses untuk mendapatkan kata dasar dengan cara menghilangkan awalan, akhiran, sisipan, dan *confixes* (kombinasi dari awalan dan akhiran)[13] dengan menggunakan library Sastrawi.

3.5 Feature Extraction

Pada tahap ini, dataset yang telah melalui tahap *preprocessing* selanjutkan akan masuk pada tahap fitur ekstraksi. Fitur ekstraksi merupakan tahap yang paling penting dalam klasifikasi, tujuan dari ekstraksi fitur itu sendiri yaitu untuk menghasilkan sekumpulan fitur dengan melakukan pembobotan kata, bobot setiap kata dihitung untuk mengklasifikasikan sebuah data [14]

TF-IDF merupakan salah satu algoritma yang paling banyak digunakan sebagai fitur ekstraksi. *Term Frequency* yaitu untuk menghitung suatu kata yang diulang dalam sebuah teks. Sedangkan Inverse Document Frequency yaitu untuk menghitung probabilitas dalam suatu kata pada teks [15]. persamaan metode TF-IDF sebagai berikut :

$$TF * IDF\ (d,t) = TF(d,t) * log\ \frac{N}{df(t)} \quad (1)$$

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.

3.6 Feature Selection

Pada tahap ini, dataset yang telah melalui tahap preprocessing dan fitur seleksi selanjutnya akan masuk pada tahap *feature selection*. Tujuan dari *feature selection* yaitu untuk mereduksi dimensi pada sebuah dataset dengan menghapus atau mengurangi fitur-fitur yang dianggap tidak diperlukan. *Feature selection* yang digunakan pada penelitian ini yaitu *Information Gain.*

Information Gain merupakan salah satu algoritma terbaik yang digunakan sebagai fitur *selection* [16]. Pada rumus 1 terlihat bahwa Y adalah kelas, X adalah atribut, $|Y_V|$ menunjukkan jumlahnya dari data dengan kelas v, dalam hal ini v terdiri dari kelas positif dan kelas negatif. |Y| adalah jumlah data di kelas positif dan kelas negatif. Entropi (Y) adalah entropi kelas Y. Entropi adalah tingkat kepentingan atribut ke kelas. Rumus Entropi (Y) dapat dilihat pada (2)

$$Entropy\ (Y) = -p_{(+)}\ log_2\ p_{(+)} - p_{(-)}\ log_2\ p_{(-)} \quad (2)$$

Y yaitu ruang atau data sample yang digunakan untuk *training*, $p_{(+)}$ yaitu probabilitas fitur yang berunilai positif pada *sample*, lalu $p_{(-)}$ yaitu probabilitas fitur yang bernilai negatif pada *sample*. Selanjutnya nilai *entropy* digunakan dalam perhitungan *gain.* jumlah kelas. Adapun persamaan 3 nilai gain dari suatu fitur [17].

$$GAIN\ (Y,X) = Entropy\ (Y) - \sum_{ugValues(X)} \frac{|Yu|}{|Y|} \times Entropy(Yu) \quad (3)$$

X yaitu atribut, maka *v* yaitu suatu nilai yang mungkin untuk atribut X, *Values(X)* yaitu himpunan nilai yang mungkin untuk atribut X, $|Yv|$ yaitu jumlah sampel untuk nilai v, $|Y|$ yaitu jumlah seluruh sampel data, dan Entropy(Yv) merupakan entropy untuk sampel yang memiliki nilai v.

3.7 Classification

Pada tahap ini, setelah melalui tahap *preprocessing*, fitur ekstraksi dan fitur seleksi selanjutnya yaitu melakukan klasifikasi pada data *training*. Pada penelitian ini metode yang digunakan untuk klasifikasi yaitu dengan menggunakan algoritma SVM. Keluaran dari tahap ini yaitu sentimen dari setiap fitur, dan hasil dari data latih akan digunakan untuk klasifikasi pada data uji.

Support Vector Machine (SVM) merupakan pendekatan klasifikasi statistik yang didasarkan pada pemaksimalan dari margin antara *instance* dan *hyper-plane* pemisahan[18]. Konsep SVM pada dasarnya adalah upaya pencarian nilai hyperline yang terbaik pemisah antara dua buah class dalam input space[6]. Dalam *hyperplane* terdapat *pattern-pattern* sebagai bagian dari anggota *patern* lain yang terdiri dari dua buah *class* yang mempunyai nilai +1 dan -1. Dalam menentukan suatu nilai pembobotan kelas positif dan negatif atau sebaliknya, dalam SVM ditentukan berdasarkan jika nilai bobot lebih dari 0 maka diklasifikasikan kedalam positif dan sebaliknya jika nilai bobot kurang dari 0 maka diklasifikasikan kedalam kelas negatif[16]. Adapun persamaan [6] untuk membentuk *hyperline* sebagai berikut:

$$w\ .\ x + b = 0 \quad (4)$$

w merupakan *vector* bobot yang tegak lurus terhadap *hyperplane*, *x* merupakan data, dan b merupakan nilai bias. Menentukan *hyperplane* terbaik yaitu dengan cara memaksimalkan jarak *margin.* Cara memaksimalkan *margin* yaitu dengan persamaan sebagai berikut:

Persamaan garis pada 2D :

$$ax + by + c = 0 \quad (5)$$

Setelah itu, persamaan garis tersebut di ubah *x* menjadi x_1 yaitu nilai fitur pertama di data x, *y* menjadi x_2, *a* menjadi w_1, *b* menjadi w_2. Sehingga menjadi :

$$w_1x_1 + w_2x_2 + c = 0 \quad (6)$$

Asumsi sekarang di dimensi d > 1, maka persamaan diatas menjadi :

$$w_1x_1 + \ldots + w_d x_d + c = 0 \quad (7)$$

Kemudian rumus diatas diperumum menjadi :

$$\sum_{i=1}^{d} w_i x_i + c = 0 \quad (8)$$

W_i merupakan nilai bobot data ke-*i*, x_i merupakan data ke-*i*, Cara lain untuk menuliskan persamaan diatas dalam notasi vector :

w_i dan x_i =

$$x = [x_i, \ldots, x_d]^T$$
$$w = [w_i, \ldots, w_d]^T \quad (9)$$

Jadi, persamaan (8) dapat ditulis menjadi :

$$g(x) = \langle w,x \rangle + c = 0 \quad (10)$$

Untuk kasus *problem binary classification* didefinisikan sebagai berikut :

$$f(x)\begin{cases} +1, \ jika\ g(x) \geq 1 \\ 1, jika\ g(x) \leq 1 \end{cases} \quad (11)$$

Cara menyatakan *margin* secara matematis, maka sederhanakan persamaan (4) sebagai berikut :

$$y(<w,x> + c) \geq 1 \quad (14)$$

Dimana y = 1 apabila x:= positif, dan y = -1 apabila x:= negatif.

Margin merupakan proyeksi orthogonal dari vector X_+ yaitu kelas positif, dan X_- merupakan kelas negatif, ke vector $\bar{w}$ maka :

$$S = <\bar{w}, X_+ - X_->$$ (15)

Kemudian, persamaan diatas disederhanakan kembali

$$\max \frac{2}{||w||} \rightarrow \min \frac{||w||}{2} \rightarrow \min \frac{1}{2}|w|^2 \quad (16)$$

Pada penelitian ini, penulis akan menggunakan beberapa model SVM seperti menggunakan kernel *linear*, *rbf*, dan *sigmoid* untuk menyelesaikan masalah linear dan non linear. Berikut merupakan beberapa fungsi kernel, yaitu :

1. Kernel Linear : $K(x_i,x) = x_i^T.x$
2. Kernel Rbf (*Radial Basic Function)* : $K(x_i,x) = exp(-\gamma|x_i - x|^2)$, $\gamma > 0$
3. Kernel Sigmoid : $K(x_i,x) = tanh(\gamma.\ x_i^T.\ x + r)^d$, $\gamma > 0$

Dimana ini *γ, r*, dan *d* merupakan parameter kernel, dan parameter C adalah penalty akibat kesalahan dalam klasifikasi untuk masing-masing kernel.

3.8 Evaluation

Pada tahap terakhir yaitu tahap evaluasi, data yang telah diklasifikasi selanjutnya dilakukan tahap evaluasi dengan menggunakan metode *confusion matrix.*

Confusion matrix adalah suatu metode yang umumnya digunakan untuk melakukan perhitungan tingkat akurasi. *Confusion matrix* memuat informasi tentang klasifikasi yang diprediksi dengan benar oleh sebuah sistem klasifikasi[19]. Hasil dari *confusion matrix* terbagi menjadi 4 yaitu *false positive, true positif, false negative,* dan *true negative.* Adapun tabel hasil dari *confusion matrix*[20] :

Tabel 5. *Confussion Matrix*

Confusion Matrix		*Predicate*	
		Positive	*Negative*
Actual	*Positive*	*TP*	*FN*
	Negative	*FP*	*TN*

Dari 4 hasil diatas, dapat dihitung evaluasi dari klasifikasi yang telah dikerjakan. Adapun rumus-rumus untuk menghitung evaluasi sebagai berikut:

- *Accuracy* yaitu rasio prediksi benar (positif dan negatif) dengan keseluruhan data.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (17)$$

- *Precision* yaitu rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif.

$$Precision = \frac{TP}{TP + FP} \quad (18)$$

- *Recall* yaitu rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (19)$$

4. Evaluasi

Pada penelitian ini terdapat 3 skenario pengujian. Skenario pertama yaitu pengujian pada proses *preprocessing*, pengujian ini bertujuan untuk mengetahui pengaruh pada proses tersebut terhadap performa analisis sentimen pada ulasan produk. Skenario kedua yaitu membandingkan penggunaan *threshold* dan

tidak menggunakan threshold pada fitur seleksi *Information Gain*, tujuan pada skenario pengujian ini yaitu untuk mengetahui performa penggunaan *threshold* pada fitur seleksi *Information Gain.* Sedangkan pada skenario terakhir yaitu melakukan pengujian pada proses klasifikasi menggunakan metode SVM, dengan menggunakan kernel *Linear*, *RBF*, dan *Sigmoid.* Tujuan dari skenario ini yaitu untuk mengetahui performansi kernel yang ada pada metode SVM.

4.1 Hasil Pengujian

4.1.1 Hasil Analisis pengaruh proses *preprocessing*

Pada skenario pertama dilakukan 3 kali pengujian pada tahap *preprocessing*, pertama yaitu pertama pada proses *Punctuation removal, Case Folding, Normalization, Data Cleaning, Tokenizing, Stopword Removal,* dan *Stmeming*. Kedua yaitu, pada proses *Punctuation removal, Case Folding, Normalization, Data Cleaning, Tokenizing, Stemming* (tanpa *Stopword*). Dan ketiga yaitu, *pada proses Punctuation removal, Case Folding, Normalization, Data Cleaning, Tokenizing, Stopword Removal* (tanpa *Stemming)*. Data yang digunakan sudah terbagi menjadi dua data yaitu data *train* sebesar 80% dan data *test* sebesar 20%, menggunakan fitur seleksi *information gain dengan* batas threshold sebesar 0.01, dan menggunakan metode SVM dengan kernel linear menggunakan nilai C sebesar 1.0. Dapat dilihat tabel penelitian sebagai berikut :

Tabel 6. Hasil Perbandingan Proses *Preprocessing*

Stopword Removal	**Stemming**	**Perfomansi**
		Akurasi
Y	Y	85.73%
-	Y	**85.89%**
Y	-	85.32%

Berdasarkan tabel hasil skenario diatas proses *Punctuation removal, Case Folding, Normalization, Data Cleaning, Tokenizing, Stemming* (tanpa *Stopword*) mendapat nilai akurasi sebesar 85.73%. Sedangkan saat percobaan kedua yaitu pada proses *Punctuation removal, Case Folding, Normalization, Data Cleaning, Tokenizing, Stopword Removal,* dan *Stmeming* mendapatkan nilai akurasi 85.73%. Pada percobaan ketiga yaitu *pada* proses *Punctuation removal, Case Folding, Normalization, Data Cleaning, Tokenizing, Stopword Removal* (tanpa *Stemming)* mendapat nilai akurasi sebesar 85.32%.

Dapat disimpulkan bahwa penggunaan proses *preprocessing* tanpa proses stopword dapat meningkatkan performansi dikarenakan pada saat proses *stopword removal* terjadi penghapusan kata-kata yang dianggap tidak penting sehingga dapat mengubah makna kalimat sebelumnya. Hal ini menyebabkan kesalahan prediksi seperti pada kalimat “tidak membuat muka berminyak” yang memiliki arti bahwa produk tersebut bagus, sedangkan saat menggunakan proses *stopword removal* dan kata “tidak” dihilangkan tersisa kata “membuat muka berminyak”, sehingga membuat perubahan kelas seharusnya kelas positif menjadi kelas negatif.

Tabel 7. Kesalahan Klasifikasi

Data	Review	Prediksi Harga	Prediksi Produk
Asli	sunscreen termahal yang pernah gue beli ini kayanya. but it's worth it sih and will definitely buy again. sukanya sama suncreen ini: - high spf - nggak meninggalkan white cast. perfectly blends into the skin - **nggak membua**t muka berminyak - very light - doesn't clog pores produk ini berhasil membuat gue jadi mau pakai sunscreen :)	-1	1
Full preprocessing	sunscreen mahal beli kaya but its worth it sih and will definetly buy again suka sunscreen high spf tinggal white cast perectly blends into the skin **muka minyak** very light doesn’t clog pores produk hasil pakai suncreen	-1	-1
Tanpa Stopword Removal	Sunscreen mahal yang pernah saya beli ini	-1	1

	kaya but it worth it sih and will definitely buy again suka sama sunscreen ini high spf tidak white cast perfectly blends into the skin **tidak buat muka minyak** very light doesn't clog pores produk ini hasil bar saya jadi mau pakai sunscreen		

4.1.2 Pengujian Pengaruh Pada Fitur Seleksi

Pada skenario kedua ini dilakukan untuk mengetahui performansi dari penggunaan seleksi fitur. Pada skenario ini menggunaan dataset terbaik dari skenario sebelumnya yaitu pada proses diatas proses *Punctuation removal*, *Case Folding*, *Normalization*, *Data Cleaning*, *Tokenizing*, *Stemming* (tanpa *Stopword*), selanjutnya menggunakan proses ekstraksi fitur dengan TF-IDF dan menggunakan metode SVM dengan kernel linear menggunakan nilai C sebesar 1.0. Dapat dilihat tabel penelitian sebagai berikut :

Tabel 8. Hasil Perbandingan Fitur Seleksi

Treshold	**Jumlah Fitur**	**Performansi**
		Akurasi
Tidak menggunakan Fitur Seleksi	**9558**	**86.14%**
0.01	3,032	85.89%
0.02	1,884	85.35%
0.03	1,318	84.94%
0.04	1,091	84.37%
0.05	916	83.52%

Berdasarkan tabel 8, hasil akurasi tertinggi didapatkan oleh percobaan pertama yaitu tidak menggunakan Fitur seleksi IG dan klasifikasi SVM mendapat nilai akurasi tertinggi sebesar 86.14%, sedangkan pecobaan dengan menggunakan fitur seleksi hasil tertinggi didapatkan oleh menggunakan batas *threshold* 0.01 dengan akurasi yang diperoleh yaitu 85.89%, sedangkan pada penelitian [5] mendapat akurasi sebesar 82,19%, hal ini kemungkinan disebabkan karena perbedaan jumlah fitur pada dataset yang digunakan pada penelitian. Adapun kesalahan pada klasifikasi dikarenakan terjadinya salah prediksi pada datatest, kemungkinan fitur yang berkontribusi dalam sentimen tidak terseleksi oleh proses IG. Kesalahan pada aspek *price* saat ketika menggunakan batas *threshold* 0,01 data yang salah sebesar 349 data, pada batas *threshold* 0,02 data yang salah sebesar 361 data, selanjutnya pada batas *threshold* 0,03 data yang salah sebesar 370, lalu selanjutnya pada batas *threshold* 0,04 data yang salah sebesar 379 data dan pada batas *threshold* 0,05 data yang salah pada baris 395 data.

Tabel 9. **Top fitur berdasarkan IG**

Terms	IG *Price*
ini	1.000000
dan	0.914802
tidak	0.890733
yang	0.813524
saya	0.808820
harga	0.736779
di	0.683292
pakai	0.683292
kulit	0.682104
jadi	0.621902

Terms	IG *Packaging*
ini	1.000000
tidak	0.979645
dan	0.956527
yang	0.855546
saya	0.825226
di	0.757408
jadi	0.680203
kulit	0.644685
pakai	0.642113
tapi	0.622740

Terms	IG *Product*
ini	1.000000
tidak	0.948158
dan	0.909686
saya	0.816054
yang	0.802552
di	0.712023
pakai	0.669345
kulit	0.664317
jadi	0.640027
tapi	0.572601

Terms	IG Aroma
ini	1.000000
tidak	0.917171
dan	0.891725
saya	0.810938
yang	0.799114
di	0.677410
pakai	0.657938
kulit	0.646956
jadi	0.591252
banget	0.577596

Pada tabel 9 merupakan top 10 fitur berdasarkan nilai IG pada masing-masing label. Kata "ini' mendapatkan nilai IG tertinggi sebesar 1,000000 pada setiap masing-masing label. Top 10 fitur ini merupakan kata yang sering muncul dan kata yang paling umum digunakan terlebih tidak menggunakan *stopword removal*. Nilai IG tersebut yang akan memasuki tahapan seleksi fitur untuk dihitung menggunakan *threshold* dan dilanjutkan pada proses klasifikasi.

Tabel 10. Perbandingan akurasi *data train* dan *data test*

	Train	Test
Tanpa IG	94.95%	86.14%
IG	89.63%	85.89%

Tabel 10 merupakan hasil perbandingan antara data train dan data test dengan menggunakan IG dan tanpa IG. Pada percobaan dengan tidak menggunakan IG nilai akurasi t*rain* memiliki selisih lebih besar daripada nilai akurasi *test*, hal tersebut menjadikan bahwa dengan tanpa menggunakan IG terjadi *overfitting*. Sedangkan dengan menggunakan IG terjadi penurunan pada akurasi *train* yang mampu mendekati akurasi *test*, dengan demikian maka dengan menggunakan IG dapat mengurangi masalah *overfitting*.

Dapat disumpulkan dari keenam percobaan, yang menghasil kan akurasi tertinggi sebesar 86.14% didapatkan oleh model dengan tidak menggunakan fitur seleksi, hal ini disebabkan karena jumlah fitur sangat mempengaruhi performansi semakin sedikit jumlah fitur maka akan semakin kecil juga nilai akurasi. Adapun total kesalahan klasifikasi pada semua aspek saat menggunakan model dengan tidak menggunakan fitur seleksi yaitu sebesar 344 data. Sedangkan saat menggunakan IG mendapatkan nilai akurasi sebesar 85.89%, total kesalahan klasifikasi pada semua aspek sebesar 349 data. Saat menggunakan IG fitur akan berkurang dibandingkan tanpa menggunakan IG, semakin banyak jumlah fitur akan sangat mempengaruhi pada hasil akurasi, sehingga hasil perolehan IG fitur yang memiliki nilai tinggi hanya akan dipilih oleh sistem. Meskipun fitur seleksi IG tidak dapat meningkatkan performansi, namun dapat mengurangi masalah komputasi. Dapat dilihat pada tabel berikut :

Tabel 11. Perbandingan Pengukuran Waktu

	Pengukuran Waktu
Tanpa IG	812.356 detik
IG	51.819 detik

Dapat dilihat dari tabel 11, bahwa dengan menggunakan fitur seleksi IG pengukuran waktu lebih cepat dibandingkan dengan tanpa menggunakan IG. Hal tersebut membuktikan bahwa dengan menggunakan fitur seleksi IG dapat mengurangi masalah komputasi.

4.1.3 Hasil Analisis pada proses klasifikasi menggunakan metode SVM, dengan menggunakan kernel *Linear, RBF,* dan *Sigmoid*

Pada skenario ketiga ini, penulis melalukan percobaan terhadap beberapa parameter yang terdapat pada SVM yaitu kernel, dan nilai C. Penulis melakukan tiga percobaan yaitu dengan menggunakan kernel linear, rbf, dan sigmoid, dengan menggunakan nilai C sebesar 1.0. Setiap percobaan ketiga kenel tersebut menggunakan fitur ekstraksi dan menggunakan batas threshold sebesar 0.01. Dapat dilihat tabel penelitian sebagai berikut :

Tabel 12. Hasil Perbandingan Klasifikasi SVM

Kernel	**Performansi**
	Akurasi
Linear	85.89%
Rbf	85.25%
Sigmoid	**85.98%**

Berdasarkan tabel tersebut, memiliki perbedaan yang tidak terlalu signifikan pada nilai akurasinya. Pada percobaan pertama dengan menggunakan kernel linear mendapat nilai akurasi sebesar 85.89%. Selanjutnya pada percobaan kedua dengan menggunakan kernel rbf mendapatkan nilai akurasi sebesar 85.25%. Pada percobaan terakhir dengan menggunakan kernel sigmoid mendapatkan nilai

akurasi sebesar 85.98%. Dapat disimpulkan dari ketiga percobaan diatas terlihat bahwa kernel terbaik adalah kernel sigmoid dengan memiliki akurasi paling tinggi, hal ini menunjukan bahwa klasifikasi review sangat tepat menggunakan algoritma SVM dengan mengggunakan kernel sigmoid.

5. Kesimpulan

Berdasarkan hasil skenario pengujian yang telah dilakukan, dapat disimpulkan bahwa hasil terbaik untuk analisis pada produk kecantikan adalah dengan menggunakan proses *Preprocessing (Punctuation removal, Case Folding, Normalization, Data Cleaning, Tokenizing, Stemming* (tanpa *Stopword*)), dengan menggunakan ekstraksi fitur *TF-IDF*, seleksi fitur *Information Gain* dengan menggunakan batas *threshold* sebesar 0,01, dan menggunakan proses klasifikasi SVM yang menghasilkan nilai akurasi sebesar 85.89%.

Proses analisis tanpa menggunakan fitur seleksi *information Gain* mendapatkan hasil akurasi lebih baik dibandingan dengan menggunakan fitur seleksi *Information Gain.* Hasil akurasi tanpa menggunakan *Information Gain* mendapatkan nilai akurasi sebesar 86.14%, sehingga dapat disimpulkan bahwa fitur seleksi menggunakan *Information Gain* tidak dapat meningkatkan performansi tetapi dapat mengurangi masalah komputasi, dikarenakan semakin banyak jumlah fitur akan mempengaruhi hasil akurasi.

Metode SVM dengan menggunakan kernel Sigmoid dapat meningkatkan performansi lebih baik dengan menghasilkan nilai akurasi yang paling tinggi diantara kernel linear dan rbf, Kernel sigmoid memiliki nilai akurasi sebesar 85.98%.

Oleh sebab itu berdasarkan penjelasan diatas bahwa fitur seleksi menggunakan *information gain* tidak dapat meningkatkan performansi, saran untuk penelitian selanjutnya yaitu untuk melakukan variasi skenario pengujian yang lebih banyak lagi untuk mendapatkan hasil performasi terbaik dari algoritma yang telah dilakukan, ataupun dengan menggunakan fitur seleksi lainnya seperti *decision tree* untuk mendapatkan hasil performansi terbaik.

REFERENSI

[1] W. Zheng and Q. Ye, "Sentiment classification of Chinese traveler reviews by support vector machine algorithm," *3rd Int. Symp. Intell. Inf. Technol. Appl. IITA 2009*, vol. 3, pp. 335–338, 2009, doi: 10.1109/IITA.2009.457.

[2] Rahul, V. Raj, and Monika, "Sentiment Analysis on Product Reviews," *Proc. - 2019 Int. Conf. Comput. Commun. Intell. Syst. ICCCIS 2019*, vol. 2019-Janua, pp. 5–9, 2019, doi: 10.1109/ICCCIS48478.2019.8974527.

[3] D. A. Kristiyanti, "Analisis Sentimen Review Produk Kosmetik Melalui," pp. 74–81, 2015.

[4] D. Y. Liliana, A. Hardianto, and M. Ridok, "Indonesian News Classification using Support Vector Machine," vol. 5, no. 9, pp. 1015–1018, 2011.

[5] S. Widya Sihwi, I. Prasetya Jati, and R. Anggrainingsih, "Twitter Sentiment Analysis of Movie Reviews Using Information Gain and Naïve Bayes Classifier," *Proc. - 2018 Int. Semin. Appl. Technol. Inf. Commun. Creat. Technol. Hum. Life, iSemantic 2018*, pp. 190–195, 2018, doi: 10.1109/ISEMANTIC.2018.8549757.

[6] P. A. Ulyah, I. Asror, and Y. R. Murti, "Analisis Sentimen untuk Review Hotel menggunakan Algoritma Support Vector Machine (SVM)."

[7] N. M. Shelke, V. Thakre, and S. Deshpande, "Identification of scope of valence shifters for sentiment analysis of product reviews," *Proc. - 2016 6th Int. Symp. Embed. Comput. Syst. Des. ISED 2016*, pp. 265–269, 2017, doi: 10.1109/ISED.2016.7977094.

[8] N. Octaviani Faomasi Daeli, "Sentiment Analysis on Movie Reviews Using Information Gain and K-Nearest Neighbor," *J. Data Sci. Its Appl.*, vol. 3, no. 1, pp. 1–007, 2020, [Online]. Available: http://commdis.telkomuniversity.ac.id/jdsa/index.php/jdsa/article/view/22.

[9] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014, doi: 10.1016/j.asej.2014.04.011.

[10] I. Khafidatul and K. Indra, "Pengaruh Ulasan Produk, Kemudahan, Kepercayaan, dan Harga Terhadap Marketplace Shopee di Mojekerto," *J. Manaj.*, vol. 6, no. 1, pp. 31–42, 2020, [Online]. Available: http://www.maker.ac.id/index.php/maker.

[11] T. Akhir, "Analisis Sentimen Level Aspek pada Ulasan Produk Kecantikan Berbahasa Indonesia Menggunakan Random Forest Program Studi Sarjana Informatika Fakultas Informatika Universitas Telkom Bandung," 2021.

[12] A. Budiarti, "Bab 2 landasan teori," *Apl. dan Anal. Lit. Fasilkom UI*, pp. 4–25, 2006.

[13] T. Kurniawan, "Implementasi Text Mining Pada Analisis Sentimen Pengguna Twitter Terhadap Media Mainstream Menggunakan Naïve Bayes Classifier Dan Support Vector Machine Media Mainstream Menggunakan Naïve Machine," *IT J.*, vol. 23, p. 1, 2017.

[14] M. S. Park and J. Y. Choi, "Theoretical analysis on feature extraction capability of class-augmented PCA," *Pattern Recognit.*, vol. 42, no. 11, pp. 2353–2362, 2009, doi: 10.1016/j.patcog.2009.04.011.

[15] S. M. H. Dadgar, M. S. Araghi, and M. M. Farahani, "A novel text mining approach based on TF-IDF and support vector machine for news classification," *Proc. 2nd IEEE Int. Conf. Eng. Technol. ICETECH 2016*, no. March, pp. 112–116, 2016, doi: 10.1109/ICETECH.2016.7569223.

[16] O. Somantri and D. Apriliani, "Support Vector Machine Berbasis Feature Selection Untuk Sentiment Analysis Kepuasan Pelanggan Terhadap Pelayanan Warung dan Restoran Kuliner Kota Tegal," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, p. 537, 2018, doi: 10.25126/jtiik.201855867.

[17] T. Akhir, "Analisis Pengaruh Seleksi Fitur Information Gain dan Mutual Information pada Klasifikasi Sentimen Ulasan Film Menggunakan Support Vector Machine Program Studi Sarjana S1 Informatika Fakultas Informatika Universitas Telkom Bandung," 2019.

[18] "Random Forest and Support Vector Machine based Hybrid Approach to SA --RF.pdf." .

[19] B. Gunawan, H. S. Pratiwi, and E. E. Pratama, "Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes," *J. Edukasi dan Penelit. Inform.*, vol. 4, no. 2, p. 113, 2018, doi: 10.26418/jp.v4i2.27526.

[20] M. Awad and R. Khanna, *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*, no. May 2016. 2015.