Perbandingan Algoritma Machine Learning untuk Analisis Sentimen Berbasis Aspek pada *Review Female Daily*

1st Muhammad Hadiyan Wicaksono
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
hadicaksono@student.telkomuniversity.ac.id

2nd Mahendra Dwifebri Purbolaksono
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
mahendradp@telkomuniversity.ac.id

3rd Said Al Faraby
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
alfaraby@telkomuniversity.ac.id

Abstrak—Beredar produk produk kecantikan yang di jual di *internet* oleh berbagai macam produsen baik luar negeri maupun dalam negeri. Akan tetapi masih diragukan kualitas kosmetik yang dijual oleh tiap produsen, agar mengetahui apakah produk tersebut baik digunakan maka produsen perlu mendapatkan ulasan/*review* dari konsumen yang memakai produk tersebut. Untuk itu agar produsen lebih mudah untuk mencari produk yang relevan dengan kesehatan maka dibutuhkan sebuah sistem untuk mengklasifikasikan *review* produk tersebut termasuk kategori relevan atau tidak relevan terhadap aspek kesehatan. Pada Tugas Akhir ini digunakan *Machine learning* pada klasifikasi sentimen menggunakan *Random Forest*, *Support Vector Machine*(SVM), dan *K-Nearest Neighbour*(KNN) untuk mencari *accuracy* tertinggi dan *F1-score* dari ketiga algoritma tersebut dengan menggunakan *feature extraction* yaitu *chi-square* dengan *feature selection* menggunakan *Selected K Best* untuk proses *preprocessing.* Dalam penelitian ini telah diperoleh analisis hasil bahwa algoritma SVM dengan kernel *Linear* mendapatkan nilai akurasi terbaik sebesar 67.10%.

Kata kunci—perbandingan, analisis sentimen, KNN, *random forest*, SVM, *chi- square, selected K Best*, *female daily*, kesehatan

Abstract—Circulating beauty products that are sold on the internet by various kinds of producers both foreign and domestic. But still the quality of the cosmetics sold by each manufacturer is doubtful, in order to know whether the product is good to use then the manufacturer needs to get reviews / reviews from consumers who use the product. For that manufacturers find it easier to find products that are relevant to health then we need a system to classify these product reviews including categories relevant or irrelevant to the health aspect. On This final project uses machine learning on sentiment classification using Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbor(KNN) to find the highest accuracy and F1-score of the three The algorithm uses feature extraction, namely chi-square with feature selection using Selected K Best for the process preprocessing. In this study, analysis of the results has been obtained that the SVM algorithm with the Linear kernel gets the best accuracy value of 67.10%.

Keywords— comparison, Sentiment Analysis, KNN, random forest, SVM, chi-square, selected K-Best, female daily, sanity

I. PENDAHULUAN

A. Latar Belakang

Beredar produk produk kecantikan yang di jual di internet oleh berbagai macam produsen baik luar negeri maupun dalam negeri. Akan tetapi masih diragukan kualitas kosmetik yang dijual oleh tiap produsen, agar mengetahui apakah produk tersebut baik digunakan maka produsen perlu mendapatkan ulasan/review dari konsumen yang memakai produk tersebut. Dengan banyaknya ulasan terhadap produk tersebut, konsumen lain dapat melihat dampak dan kegunaan produk tersebut sesuai kebutuhan konsumen itu sendiri. Terutama di bidang kesehatan, apakah produk yang di pakai memiliki dampak pada kulit, seperti jerawat, bibir kering, kulit terbakar, dan masalah kulit lainnya. Salah satu forum yang menampung review tentang produk kecantikan adalah Female Daily.Di dalam forum tersebut produsen dapat mengetahui apa kekurangan dari produk yang terdapat di dalam produk tersebut dengan menerima review dari para pengguna produk tersebut terutama review di bidang kesehatan.

Untuk itu agar produsen lebih mudah untuk mencari produk yang relevan dengan kesehatan maka dibutuhkan sebuah sistem untuk mengklasifikasikan review produk tersebut termasuk kategori relevan atau tidak relevan terhadap aspek kesehatan.

Pada Tugas Akhir ini digunakan Machine learning pada klasifikasi sentimen menggunakan Random Forest, Support Vector Machine(SVM), dan K-Nearest Neighbour(KNN) untuk mencari akurasi tertinggi dan *F1- score* dari ketiga algoritma tersebut dengan menggunakan feature extraction yaitu chi-square dengan *feature selection* menggunakan *Selected K Best* untuk proses *preprocessing*.

B. Topik dan Batasannya

Dengan adanya permasalahan seperti yang telah dijelaskan pada latar belakang maka dapat disimpulkan oleh penulis permasalahan yang dapat disimpulkan adalah dengan mengetahui cara untuk mengklasifikasikan data teks berdasarkan data *review Female Daily* yang telah ada, dan juga menganalisis hasil teks dari data teks *review*

Female Daily yang telah melalui proses klasifikasi.

C. Tujuan

Tujuan pada penelitian ini sesuai yang sudah dijelaskan pada latar belakang adalah peneliti dapat menerapkan metode klasifikasi SVM, KNN, *Random Forest* menggunakan *chi-square* dalam analisis sentimen berbasis aspek kesehatan dalam review produk *Female Daily,* dan dapat melakukan analisis dari hasil klasifikasi perbandingan tiga algoritma tersebut dengan menggunakan seleksi fitur TF-IDF

D. Organisasi Tulisan

Penulisan tugas akhir ini tersusun dalam beberapa bagian, yaitu sebagai berikut:

1. Pendahuluan
   Menjelaskan latar belakang, rumusan masalah dan tujuan dari topik yang diambil
2. Studi Terkait
   Berisi penelitian-penelitian sebelumnya yang digunakan sebagai literatur acuan dalam pembuatan tugas akhir ini
3. Sistem yang dibangun
   Menggambarkan alur berjalannya system yang berawal dari metodologi penelitian untuk mencapai objektif, rancangan analisis hingga metode pengujian yang dilakukan
4. Evaluasi
   Analisis terhadap sentimen yang digunakan menjelaskan hasil klasifikasi algoritma yang digunakan menggunakan seleksi fitur TF-IDF
5. Kesimpulan
   Penelitian yang dilakukan serta saran untuk penelitian selanjutnya

II. KAJIAN TEORI

Dalam penelitian ini penulis mendapat beberapa referensi berupa paper atau jurnal dari penelitian sebelumnya yaitu:

Pada penelitian [1] Ekky Yulianti Prastika, dengan judul "Analisis Sentimen pada Ulasan Produk Kecantikan Menggunakan K-Nearest Neighbor dan Information Gain" disimpulkan bahwa data yang tidak seimbang mendapatkan akurasi paling tinggi karena model cenderung memprediksi label mayoritas. Pada tahapan preprocessing, penerapan stemming dan normalisasi terbukti memberikan pengaruh yang signifikan karena dapat mengurangi kata yang memiliki satu makna tetapi mempunyai lebih dari satu kata atau term. Kemudian, pada seleksi fitur, nilai IG lebih dari 0,5 mempunyai tingkat relevansi yang tinggi terhadap kelasnya. Nilai k yang optimal diperoleh sebesar 23. Sehingga, dari ketiga skenario yang dilakukan, akurasi tertinggi diperoleh oleh dataset yang menerapkan stemming, normalisasi, IG dengan threshold 0,5, dan k = 23 dengan akurasi sebesar 74,21%.

Penelitian yang dilakukan oleh Novelty Octaviani Faomasi Daeli [2] 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

Penelitian yang dilakukan oleh Willy Wildan Kamal [3] pada tahun 2021 dengan judul "Analisis Sentimen Ulasan Produk Skincare Menggunakan Metode Support Vector Machine" bahwa pada penelitian ini, Dataset terdiri dari 1260 ulasan terbagi menjadi 628 ulasan positif, 174 ulasan netral, dan 458 ulasan negatif. Klasifikasi dengan metode SVM dengan kernel linear diperoleh tingkat akurasi sebesar 86.9 %, akan tetapi pada penelitian ini memiliki masalah yaitu, Dataset yang digunakan baik jumlah data positif, jumlah data netral, maupun jumlah data negatif tidak memiliki perbandingan yang sama, dan Sistem belum mampu membedakan kata-kata antara kondisi usersaat ini dengan hasil setelah pemakaian produk, sehingga masih ada kesalahan dalam prediksi.

Pada penelitian [4] yang dilakukan oleh Muhammad Asjad Adna Jihad dengan judul "Analisis Sentimen Terhadap Ulasan Film Menggunakan Algoritma Random Forest" pada tahun 2021 dapat disimpulkan bahwa dari kedua skenario pengujian, pertama, banyak dimensi *skip-gram word2vec* dapat mempengaruhi hasil akhir performansi namun tidak berarti semakin besar jumlah dimensi yang digunakan maka semakin baik hasil performansi yang didapatkan. Kedua, terbukti bahwa penerapan metode *stemming* terhadap *dataset* dapat mempengaruhi hasil akhir performansi dimana pada kasus kali ini *stemming* meningkatkan performansi lebih jauh.

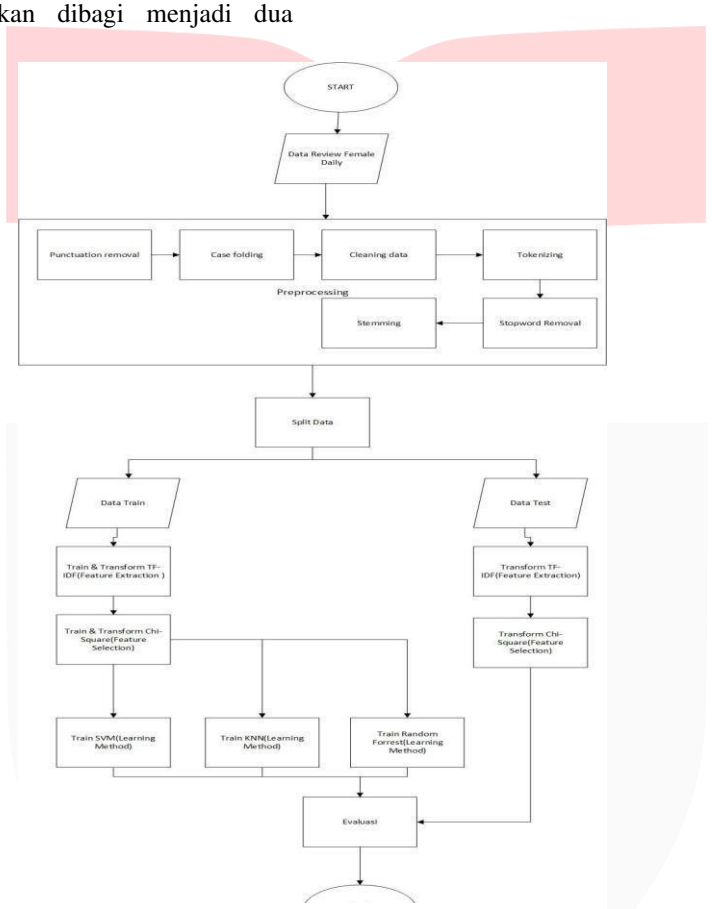
Pada penelitian [5] Riszki Wijayatun Pratiwi, dengan judul "Analisis Sentimen Pada Review Skincare Female Daily Menggunakan Metode Support Vector Machine (SVM)" pada tahun 2021, dapat disimpulkan bahwa Metode Support Vector Machine dapat melakukan prediksi kelas sentimen pada review produk kecantikan sesuai analisa yang disiapkan. Kemudian analisis sentimen ini belum sepenuhnya relevan untuk memprediksi kelas sentimen yang sesuai terhadap pemberian kelas berdasarkan rating atau bintang pada review tersebut, dan analisis sentimen pada produk kecantikan menggunakan SVM menghasilkan nilai akurasi menggunakan dataset 80% data training dan 20% data testing

sehingga mendapatkan akurasi sebesar 87% dengan recall sebesar 90%, precision sebesar 84,90%, dan f1 score sebesar 87,37%

III. METODE

A. Skema umum

Pada penelitian ini terdapat skema umum pengujian dimana data yang telah diperoleh dari ulasan/*review website Female Daily* akan melalui proses *preprocessing* dimana data akan melalui proses *Punctional removal*, *Case Folding*, *Cleaning data, Tokenizing*, *Stopword removal*, dan *Stemming*. Lalu setelah tahapan proses *preprocessing* data akan dibagi menjadi dua bagian yaitu data *train* dan data *test*. Setelah tahapan pemisahan data menjadi 2 bagian akan dilakukan proses *feature extraction transform TF-IDF* untuk menghitung kata yang berulang dalam teks, setelah melakukan proses *feature extraction* akan dilakukan tahapan selanjutnya yaitu proses *feature selection* dengan menggunakan *transform chi-square*. Data-data hasil dari proses *feature selection* akan di proses menggunakan tiga algoritma klasifikasi yaitu SVM, KNN, serta *Random Forest*. Hasil klasifikasi dari tiga algoritma tersebut akan dilakukan proses analisis untuk mendapatkan algoritma terbaik. Berikut merupakan gambar skema umum penelitian:



GAMBAR 1
SKEMA UMUM

B. Dataset

Dataset yang digunakan dalam penelitian ini diambil dari beberapa review atau ulasan produk yang terdapat pada situs Female Daily. Review atau ulasan yang digunakan di penelitian ini adalah ulasan yang berupa produk kecantikan dengan yang mengacu tentang kesehatan wajah dan kulit, dengan menggunakan Bahasa Indonesia dan Bahasa inggris namun lebih banyak memakai Bahasa Indonesia. Total ulasan yang digunakan dalam penelitian adalah sebanyak 15000 data dan memiliki 4 atribut yaitu *id_review, clean_review*, *sentiment*, *contain_health*, dan pada penelitian ini dilakukan penelitian pada atribut *contain_health* yang telah di beri label 0 untuk label tidak relevan, dan 1 untuk label relevan.

Berikut merupakan dua contoh ulasan produk kecantikan dari dataset penelitian dengan
contain_health bernilai 1 yaitu relevan dengan kesehatan dan 0 yautu tidak relevan:

TABLE 1

BEBERAPA DATASET YANG DIGUNAKAN

Id_review	clean_review	contain_health
129	selalu perlu pinkish nude lipstick. Saat tahu ttg tom ford lipstick, dan cocok shadenya, langsung beli. Saya termasuk yang 7ensitive dengan lipstick, gampang kering bibirnya, jadi saya jarang beli long lasting lipstick. Liptick ini tidak long lasting, tapi ternyata bertahan cukup lama, selesai breakfast, masih ada lho. Formulanya lembut, gampang diaplikasikan dan pigmented. Yang pasti tidak membuat bibir saya kering	1
130738	suka banget sama lip ice sheer color, dari sma sampe sekarang kuliah masih pake tiap hari karena aku ga suka pake lipstick. pake lipstick paling kalo ada acara2 formal aja. ya ini solusi yg tepat buat orang kayak aku yg ga suka pake lipstick tiap hari, namun bibir tetap berwarna pink natural dan ga kering atau pecah-pecah :)	0

Pada penelitian ini *class contain_health* diberikan label dan pendistribusian label yang dipakai adalah:

TABLE 2.
JUMLAH MASING-MASING LABEL PADA ASPEK CONTAIN HEALTH

Aspek	Sentimen		Total Data Setiap Aspek
	1	0	
contain_health	6335(42.23%)	8665(57.77)	15000

C. Preprocessing

Dataset yang telah dipersiapkan harus melalui alur proses preprocessing terlebih dahaulu yang terdiri dari:

1. Punctuation Removal
   Dalam Proses ini terjadi proses pembersihan data dengan menghapus tanda baca, angka, atau karakter khusus selain kata dari dataset yang digunakan
2. Case Folding
   Dalam proses huruf huruf pada dataset akan diubah menjadi huruf kecil(lower case)
3. Cleaning data
   Dalam proses ini terjadi proses penghapusan tanda atau simbol dan juga menghapus angka
4. Tokenizing
   Dalam proses ini terjadi proses pemisahan kata-kata menjadi potongan-potongan yang disebut sebagai token yang nantinya dapat di analisa
5. Stopword Removal
   Dalam proses ini terjadi penghapusan kata-kata yang berpengaruh damenentukan suatu kategori sentimen. Dalam penelitian ini menggunakan list *stopword removall* yaitu *indonesian*
6. Stemming
   Dalam proses ini terjadi proses untuk mendapatkan kata dasar dengan cara
   menghilangkan awalan, akhiran, kata sisipan, dan *confixes* (kombinasi kata awalan dan kata akhiran dengan menggunakan library Sastrawi untuk kata kata dalam bahasa Indonesia

D. Split Data

Pada proses ini dilakukan pembagian data, data dibagi menjadi 2 yaitu data train dan data test dimana data training berjumlah 80% dari dataset sementara data test berjumlah 20% dari dataset. Pada proses split data ini digunakan random state untuk membuat nilai konsisten saat system dijalankan

E. Feature Extraction

Dalam proses ini dataset yang telah melalui tahap *preprocessing* lalu dilanjutkan dengan proses split data akan dilanjutkan tahapan fitur ekstraksi. Fitur ekstraksi merupakan tahap yang paling penting sebelum proses klasifikasi tujuan dari tahapan ini adalah untuk mengahasilkan sekumpulan fitur dengan melakukan pembobotan kata, pembobotan kata tersebut dapat dihitung sehingga dapat mengklasifikasikan sebuah data

TF-IDF merupakan metode yang digunakan untuk memberikan bobot pada *term* sebagai strategi untuk mengklasifikasikan dokumen. Proses pembobotan pada *term* menggunakan TF-IDF terdiri dari menghitung nilai TF(Term Frequency) yaitu untuk menghitung suatu kata yang diulang dalam sebuah teks, sementara IDF(Invers Document Frequency) yaitu untuk menghitung probabilitas dalam suatu kata pada teks. Berikut merupakan perhitungan pembobotan yang digunakan dalam metode TF-IDF:

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

Keterangan:
TF : Merupakan nilai dari *term frequency*
IDF(d,t) : Merupakan nilai dari *inverse document frequency*
(d,t) : Merupakan banyaknya term t pada dokumen d N : Jumlah dari semua kumpulan dokumen
df(t): Merupakan jumlah dokumen yang memiliki term t

F. Feature Selection

Dalam proses ini dataset yang telah melalui tahap *feature extraction* menggunakan TF-IDF, akan dilakukan proses fitur seleksi dengan menggunakan metode Chi-Square

Chi-square merupakan algoritma seleksi fitur untuk mengatur kurangnya fleksibilitas antara kategori dan term. Chi -square merupakan salah satu fitur seleksi yang banyak digunakan pada penelitian sebelumnya dan terbukti dapat meningkatkan akurasidalam klasifikasi penelitian sebelumnya.Chi-square memiliki persamaan sebagai berikut:

$$X^2 = \sum_{i=1}^{n} \frac{(Oi - Ei)^2}{Ei} \quad (2)$$

Keterangan:
Oi : suatu nilai di dalam i Ei : ekspektasi nilai i

G. Classification

Dalam proses ini dataset yang telah melakukan fitur ektraksi dan fitur seleksi akan diklasifikasikan menggunakan 3 metode algoritma yaitu *Support Vector Machine*(SVM), *K-Nearest Neighbors*(KNN), *Random Forest*(RF). Keluaran dari tahap ini yaitu sentimen dari setiap fitur dan hasil dari pengujian data latih/*test* akan digunakan untuk klasifikasi data uji/*train*, lalu dilakukan analisis agar dapat mengetahui klasifikasi yang lebih efisien

1. K-Nearest Neighbors(KNN)

K-Nearest Neighbors(KNN) KNN merupakan algoritma sederhana dalam machine learning yang berbasis jarak. Dalam proses KNN untuk menghitung suatu jarak terdapat suatu proses yaitu dengan *Euclidean Distance,* dengan cara menghitung jarak dari satu data yang berasal dari data uji/*test* dengan seluruh data dari data latih/*train*. Untuk menghitung jarak dapat menggunakan persamaan *Euclidean Distance*: sebagai berikut:

$$d_{euclid} = \sqrt{\sum_{i=1} |P_i - Q_i|^2} \quad (3)$$

Keterangan:
d_{euclid} : jumlah fitur atau dimensi
P_i : fitur ke I pada data uji
Q_i : fitur ke I pada data latih

2. Random Forest(RF)

Klasifikasi kedua adalah dengan menggunakan *Random Forest* (RF). *Random Forest* memiliki konsep menggabungkan beberapa *Decision Tree* untuk melakukan klasifikasi dimana setiap tree akan diberikan data secara acak dari sebagian *dataset training*

3. Support Vector Machine(SVM)

Klasifikasi ketiga adalah dengan menggunakan *Support Vector Machine* (SVM). SVM digunakan untuk mencari *hyperplane* yang optimal dengan cara memaksimalkan jarak antar kelas. *Hyperplane* adalah fungsi yang dapat dipakai memisahkan antar kelas. Berdasarkan penelitian yang dilakukan rumus SVM yang dipakai adalah sebagai berikut:

a. Kernel Linear
b. Kernel RBF
c. Kernel Sigmoid

IV. HASIL DAN PEMBAHASAN

A. Hasil Pengujian

1. Hasil Skenario pada preprocessing

Pada Skenario pertama dilakukan 3 kali pengujian dalam tahapan preprocessing. Pengujian pertama menggunakan *Punctional Removal*, *Case Folding*, *Cleaning data, Tokenizing*, *Stopword Removal*, *Stemming*. Pengujian kedua menggunakan *Punctional Removal*, *Case Folding*, *Cleaning data, Tokenizing*, *Stemming*(tanpa *Stopword)*. Pengujian ketiga menggunakan *Punctional Removal*, *Case Folding*, *Cleaning data, Tokenizing*, *Stopword Removal*(tanpa *Stemming*). Data yang digunakan terbagi menjadi dua bagian yaitu 80% untuk data *train* dan 20% untuk data *test* menggunakan fitur seleksi TF-IDF dan dengan menggunakan 3 algoritma yaitu SVM , KNN, *Random Forest*. Dengan hasil sebagai berikut:

a. SVM kernel linear C=1.00

Berikut merupakan tabel hasil klasifikasi dari perbandingan proses *preprocessing*, menggunakan *stopword removal* tanpa *stemming,* dengan *stemming* tanpa *stopword removal*, dan dengan keduanya yaitu *stemming* dan *stopword,* dengan algoritma klasifikasi menggunakan SVM dengan kernel *linear* dengan nilai C sebesar 1.00:

TABLE 3
HASIL PERBANDINGAN PROSES PREPROCESSING SVM KERNEL LINEAR

Stopword Removal	**Stemming**	**Performansi**
		Akurasi
Y	Y	67.10%
-	Y	63.70%
Y	-	67.10%

Berdasarkan tabel hasil skenario diatas pada proses *Punctional Removal*, *Case Folding*, *Cleaning data, Tokenizing*, *Stopword Removal*, *Stemming*, memiliki akurasi sebesar 67.1%
.Sedangkan saat percobaan kedua yaitu pada proses *Punctional Removal*, *Case Folding*, *Cleaning data, Tokenizing*, *Stemming*(tanpa *Stopword)* memiliki akurasi sebesar 63.7%.Pada percobaan terakhir menggunakan proses *Punctional Removal*, *Case Folding*, *Cleaning data, Tokenizing*, *Stopword Removal*(tanpa *Stemming*) akurasi yang didapatkan sebesar 67.1% .

b. kernel sigmoid C=1.00

Berikut merupakan tabel hasil klasifikasi dari perbandingan proses *preprocessing*, menggunakan *stopword removal* tanpa *stemming,* dengan *stemming* tanpa *stopword removal*, dan dengan keduanya yaitu *stemming* dan *stopword,* dengan algoritma klasifikasi menggunakan SVM dengan kernel *sigmoid* dengan nilai C sebesar 1.00:

TABLE 4
HASIL PERBANDINGAN PROSES PREPROCESSING SVM KERNEL SIGMOID

Stopword Removal	Stemming	Performansi
		Akurasi
Y	Y	64.73%
-	Y	62.80%
Y	-	64.73%

Berdasarkan tabel hasil skenario diatas pada proses *Punctional Removal*, *Case Folding*, *Cleaning data, Tokenizing*, *Stopword Removal*, *Stemming*, memiliki akurasi sebesar 64.73%
.Sedangkan saat percobaan kedua yaitu pada proses *Punctional Removal*, *Case Folding*, *Cleaning data, Tokenizing*, *Stemming*(tanpa *Stopword)* memiliki akurasi sebesar 62.8% .Pada percobaan terakhir menggunakan proses *Punctional Removal*, *Case Folding*, *Cleaning data, Tokenizing*, *Stopword Removal*(tanpa *Stemming*) akurasi yang didapatkan sebesar 64.73% .

c. kernel RBF C=1.00

Berikut merupakan tabel hasil klasifikasi dari perbandingan proses *preprocessing*, menggunakan *stopword removal* tanpa *stemming,* dengan *stemming* tanpa *stopword removal*, dan dengan keduanya yaitu *stemming* dan *stopword,* dengan algoritma klasifikasi menggunakan SVM dengan kernel RBF dengan nilai C sebesar 1.00:

TABLE 5
HASIL PERBANDINGAN PROSES PREPROCESSING SVM KERNEL RBF

Stopword Removal	Stemming	Performansi
		Akurasi
Y	Y	62.43%
-	Y	62.80%
Y	-	62.43%

Berdasarkan tabel hasil skenario diatas pada proses *Punctional Removal*, *Case Folding*, *Cleaning data, Tokenizing*, *Stopword Removal*, *Stemming*, memiliki akurasi sebesar 62.43%
.Sedangkan saat percobaan kedua yaitu pada proses *Punctional Removal*, *Case Folding*, *Cleaning data, Tokenizing*, *Stemming*(tanpa *Stopword)* memiliki akurasi sebesar 62.8% .Pada percobaan terakhir menggunakan proses *Punctional Removal*, *Case Folding*, *Cleaning data, Tokenizing*, *Stopword Removal*(tanpa *Stemming*) akurasi yang didapatkan sebesar 62.43% .

B. KNN K=50

Berikut merupakan tabel hasil klasifikasi dari perbandingan proses *preprocessing*, menggunakan *stopword removal* tanpa *stemming,* dengan *stemming* tanpa *stopword removal*, dan dengan keduanya yaitu *stemming* dan *stopword,* dengan algoritma klasifikasi menggunakan KNN(K-Nearest Neighbors) dengan nilai K sebesar 50:

TABLE 6
HASIL PERBANDINGAN PROSES PREPROCESSING KNN

Stopword Removal	Stemming	Performansi
		Akurasi
Y	Y	60.00%
-	Y	59.96%
Y	-	60.00%

Berdasarkan tabel hasil skenario diatas pada proses *Punctional Removal*, *Case Folding*, *Cleaning data, Tokenizing*, *Stopword Removal*, *Stemming*, memiliki akurasi sebesar 60% .Sedangkan saat percobaan kedua yaitu pada proses *Punctional Removal*, *Case Folding*, *Cleaning data, Tokenizing*, *Stemming*(tanpa *Stopword)* memiliki akurasi sebesar 59.96% .Pada percobaan terakhir menggunakan proses *Punctional Removal*, *Case Folding*, *Cleaning data, Tokenizing*, *Stopword Removal*(tanpa *Stemming*) akurasi yang didapatkan sebesar 60% .

1. Random Forest

Berikut merupakan tabel hasil klasifikasi dari perbandingan proses *preprocessing*, menggunakan *stopword removal* tanpa *stemming,* dengan *stemming* tanpa *stopword removal*, dan dengan keduanya yaitu *stemming* dan *stopword,* dengan algoritma klasifikasi menggunakan *Random Forest* dengan nilai *random_state* sebesar 0 :

TABLE 7
HASIL PERBANDINGAN PROSES PREPROCESSING RANDOM FOREST

Stopword Removal	**Stemming**	**Performansi**
		Akurasi
Y	Y	65.76%
-	Y	60.43%
Y	-	65.76%

Berdasarkan tabel hasil skenario diatas pada proses *Punctional Removal*, *Case Folding*, *Cleaning data, Tokenizing*, *Stopword Removal*, *Stemming*, memiliki akurasi sebesar 65.76%
.Sedangkan saat percobaan kedua yaitu pada proses *Punctional Removal*, *Case Folding*, *Cleaning data, Tokenizing*, *Stemming*(tanpa *Stopword)* memiliki akurasi sebesar 60.43% .Pada percobaan terakhir menggunakan proses *Punctional Removal*, *Case Folding*, *Cleaning data, Tokenizing*, *Stopword Removal*(tanpa *Stemming*) akurasi yang didapatkan sebesar 65.76% .

Seperti pada daftar tabel - tabel yang telah di peroleh dapat disimpulkan bahwa akurasi yang diperoleh pada *preprocessing* dalam proses *stemming* tidak mendapatkan peningkatan akurasi dikarenakan data yang diperoleh banyak menggunakan kata kata yang tidak baku dan menggunakan bahasa asing dan library yang digunakan menggunakan Sastrawi untuk pemrosesan Bahasa Indonesia seperti pada tabel berikut hasil sebelum dan setelah *stemming*:

TABLE 8
HASIL KESALAHAN KLASIFIKASI

Data	Review	Contain Health
Asli	suka banget sama lip ice sheer color, dari sma sampe sekarang kuliah masih pake tiap hari karena aku ga suka pake lipstick. pake lipstick paling kalo ada acara2 formal aja. ya ini solusi yg tepat buat orang kayak aku yg ga suka pake lipstick tiap hari, namun bibir tetap berwarna pink natural dan ga kering atau pecah-pecah :)	0
Full Preprocessing	suka banget lip ice sheer color sma sampe kuliah pake ga suka pake lipstick pake lipstick kalo acara formal aja ya solusi yg orang kayak yg ga suka pake lipstick bibir warna pink natural ga kering pecah pecah;0	0

Tanpa Stemming	suka banget lip ice sheer color sma sampe kuliah pake ga suka pake lipstick pake lipstick kalo acara formal aja ya solusi yg orang kayak yg ga suka pake lipstick bibir warna pink natural ga kering pecah pecah;0	0

A. Analisis Hasil Pengujian Perbandingan Algoritma

Pada tahapan ini penulis melakukan percobaan menggunakan algoritma KNN, Random Forest ,dan SVM dengan masing masing parameter dari algoritma tersebut. Penulis melakukan tiga percobaan yaitu dengan menggunakan KNN dengan nilai K sebesar 50, SVM menggunakan tiga kernel yaitu *RBF*, *Linear*, dan *Sigmoid* masing-masing kernel menggunakan nilai C sebesar 1.00, dan menggunakan *Random Forest*. Dapat dilihat pada tabel penelitian berikut:

TABLE 9
HASIL PERCOBAAN TIGA ALGORITMA

Algoritma		**Akurasi**
KNN		60.00%
SVM	**Linear**	**67.10%**
	Sigmoid	64.73%
	RBF	62.43%
Random Forest		65.76%

Berdasarkan tabel tersebut memiliki perbedaan yang tidak terlalu signifikan pada nilai akurasi. Pada percobaan ini menggunakan tiga algoritma klasifikasi.

Pada percobaan algoritma pertama menggunakan KNN dengan K sebesar 50 mendapat nilai akurasi sebesar 60.00%.

Selanjutnya Pada percobaan algoritma kedua yaitu SVM menggunakan tiga percobaan dengan tiga kernel pada SVM yaitu *Linear, RBF, Sigmoid*. Pertama dengan kernel *Linear* mendapatkan nilai akurasi sebesar 67.10%.Kedua menggunakan kernel *Sigmoid* mendapatkan nilai akurasi sebesar 64.73%. menggunakan kernel *RBF* mendapatkan nilai akurasi sebesar 62.43%.

Pada percobaan terakhir menggunakan algoritma *Random Forest* mendapatkan nilai akurasi sebesar 65.76%.

Dapat disimpulkan dari ketiga algoritma diatas terlihat bahwa algoritma terbaik adalah algoritma SVM dengan kernel *Linear* dengan memiliki akurasi paling tinggi yaitu sebesar 67.10%, hal ini menunjukan bahwa klasifikasi review sangat tepat menggunakan algoritma SVM dengan kernel *Linear.*

V. KESIMPULAN

Berdasarkan hasil analisis dari scenario yang telah dilakukan, maka dapat diperoleh kesimpulan yaitu:

1. Metode Stemming dapat mempengaruhi akurasi
2. Algoritma SVM dengan kernel *Linear* merupakan algoritma terbaik dengan akurasi tertinggi dengan nilai 67.10%

A. Saran

Saran yang diberikan untuk penelitian selanjutnya adalah:

1. Gunakan dua metode stemming yaitu dalam Bahasa Indonesia dan Bahasa asing
2. Dapat menambah kernel lain dalam klasifikasi SVM

REFERENSI

[1] Analisis Sentimen pada produk kecantikan dari Ulasan Female Daily Menggunakan Information Gain dan SVM Classifier Nadifa Fadila Putri ,Said AlFaraby S.T.,M.Sc., Mahendra Dwifebri P.,S.Kom.,M.kom 2021- TF-IDF

[2] Sentiment Analysisi of Beauty Product Reviews Using the K-Nearest Neighbour(KNN) and TF-IDF Methods with Chi-Square Feature Selection Yusrifa Deta Kirana, Said Al Faraby, Telkom University 2021 Bandung -Chi-square

[3] Analisis Sentimen pada Ulasan Produk Kecantikan Menggunakan *K-Nearest Neighbor* dan *Information Gain* Ekky Yulianti Prastika, Said Al Faraby, Mahendra Dwifebri P. Telkom University , Bandung, 2021 – KNN

[4] Analisis Sentimen Terhadap Ulasan Film Menggunakan Algoritma Random Forest, Muhammad Asjad Adna Jihad, Adiwijaya , Widi Astutii, Telkom University , Bandung, 2021 – Random Forest

[5] E. Y. P. S, Said Al Faraby, Mahendra Dwifebri P, "Analisis Sentimen pada Ulasan Produk Kecantikan Menggunakan K-Nearest Neighbor dan Information Gain", e-Proceeding of Engineering: Vol.8, pages 10091- 10105,2021.

[6] 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].

[7] Willy Wildan Kamal, "Analisis Sentimen Ulasan Produk Skincare Menggunakan Metode Support Vector Machine", https://dspace.uii.ac.id/handle/123456789/34153, 2021.

[8] Muhammad Asjad Adna Jihad, Adiwijaya, Widi Astuti, "Analisis Sentimen Terhadap Ulasan Film Menggunakan Algoritma Random Forest",e- Proceeding of Engineering : Vol.8 pages 10153- 10165,2021.

[9] Riszki Wijayatun Pratiwi, S. F. H., Dairoh, D. I. Af'idah, Q. R. A, A. G. F., "Analisis Sentimen Pada Review Skincare Female Daily Menggunakan Metode Support Vector Machine (SVM)", J. OF INISTA, VOL. 1, NO. 1, PP.040-046, NOV 2021.