Analisis Sentimen Ulasan Restoran Menggunakan Metode *Support Vector Machine*

Yoel Julianto, Djoni Haryadi Setiabudi, Silvia Rostianingsih
Program Studi Informatika, Universitas Kristen Petra
Jl. Siwalankerto 121 – 131 Surabaya 60236
Telp. (031) – 2983455, Fax. (031) - 8417658
yoel.julianto72@gmail.com, djonihs@petra.ac.id, silvia@petra.ac.id

ABSTRAK

Ulasan restoran yang ada di internet memiliki pengaruh yang sangat besar bagi suatu restoran. Ulasan yang diberikan membantu pelanggan lain untuk mengevaluasi usaha atau servis yang diberikan dari suatu restoran. Pelanggan dapat memberikan ulasan positif atau negatif. Banyaknya ulasan dari pelanggan membuat restoran kesulitan untuk mengetahui apakah restoran mereka memiliki lebih banyak ulasan positif atau negatif. Pada skripsi ini dibuat aplikasi untuk mengetahui restoran memiliki ulasan positif atau negatif.

Aplikasi yang dilengkapi dengan fitur *text mining* akan membantu restoran dalam mengevaluasi restorannya. Langkah yang dilakukan adalah *preprocessing* yang terdiri dari *case folding*, *tokenization*, *stopword removal*, dan *stemming*. Lalu proses mengubah data teks menjadi vektor menggunakan TF-IDF. Selanjutnya data dilatih menggunakan metode *Support Vector Machine* yang menghasilkan model yang akan digunakan untuk melakukan prediksi dari data *input*. Data yang dilatih adalah ulasan berbahasa Indonesia dari berbagai restoran.

Hasil dari penelitian yang dilakukan menghasilkan akurasi sebesar 93% dan f1-score sebear 93%. Dari penelitian ini juga menunjukkan bahwa untuk meningkatkan nilai akurasi dan f1-score, model klasifikasi SVM membutuhkan parameter TF-IDF min_df sebesar 0.05, max_df sebesar 0.75, norm l2, n-gram (1, 2), *kernel* SVM linear dengan C sebesar 1. Selain parameter TF-IDF dan SVM, jumlah dataset juga dapat meningkatkan nilai akurasi dan f1-score.

Kata Kunci: SVM, ulasan, restoran

ABSTRACT

Reviews on restaurants on the internet have a huge impact on a restaurant. Reviews provided help other customers to evaluate the business or services provided from a restaurant. Customers can leave positive or negative reviews. The large number of reviews from customers makes it difficult for restaurants to know if their restaurant has move positive or negative reviews. In this undergraduate thesis an application will be made to determine whether a restaurant has positive or negative reviews.

Application that is equipped with text mining features will help restaurant in evaluate their restaurant. The steps taken are preprocessing which consist of case folding, tokenization, stopword removal, and stemming. Then the process of converting text data into vector using TF-IDF. Furthermore the data will be trained using Support Vector Machine which later will generate a model that will be used to make predictions from input data. The data which be used as training are Indonesian-language reviews from various restaurants.

From this research conducted the result showed an accuracy of 93% and f1-score of 93%. To increase accuracy and f1-score values, classification model require TF-IDF parameters min_df 0.05, max_df 0.75, norm l2, n-gram (1, 2), linear SVM kernel with C 1. Besides TF-IDF and SVM parameters, the number of datasets can also increase confusion matrix and f1-score values.

Keywords: *SVM, review, restaurant.*

1. PENDAHULUAN

Pada era saat ini ulasan – ulasan yang ada di internet mengenai suatu restoran mempunyai pengaruh yang sangat besar bagi suatu restoran. Ulasan – ulasan yang diberikan pelanggan pada suatu restoran dapat memberikan ulasan positif atau negatif mengenai restoran tersebut. [2] *review* dan rating yang diberikan membantu pelanggan lain untuk mengevaluasi usaha atau servis yang diberikan dan mereka dapat menentukan pilihan. Dari hasil ulasan – ulasan tersebut, tidak jarang ulasan yang diberikan berisi informasi yang kurang lengkap dan tidak sesuai. Dengan banyaknya ulasan yang ada, membuat suatu restoran kesulitan untuk mengetahui restoran tersebut memiliki lebih banyak ulasan positif atau negatif. Hal seperti ini menjadi hal yang penting untuk dikaji sebagai pemrosesan teks. Untuk menyaring ulasan – ulasan yang ada, analisis sentimen sangat diperlukan. Analisis sentimen pada ulasan dilakukan untuk mengetahui ulasan bersifat positif atau negatif. [20] klasifikasi sentimen bertujuan untuk mengatasi masalah ini dengan cara otomatis mengklasifikasikan ulasan pengguna menjadi pendapat positif atau negatif.

Sudah ada beberapa penelitian sebelumnya mengenai ulasan restoran dengan menggunakan bahasa Indonesia diantaranya, yang pertama penelitian yang dilakukan oleh [9] dengan judul analisis sentimen pada *review* restoran dengan teks bahasa Indonesia menggunakan algoritma *Naïve Bayes*. Selanjutnya, ada juga penelitian yang dilakukan oleh [13] dengan judul analisis sentimen menggunakan algoritma *Naïve Bayes* terhadap *review* restoran di Singapura. Penelitian yang dilakukan oleh [3] dengan judul komparasi algoritma *Support Vector Machine* dan *Naïve Bayes* dengan algoritma genetika pada analisis sentimen calon gubernur Jabar 2018 - 2023.

Penelitian yang dilakukan [9] menggunakan penggabungan metode pemilihan fitur *Genetic Algorithm* dengan menggunakan fitur 3-gram (3 suku kata) menghasilkan peningkatan akurasi pada klasifikasi yang mencapai 4% dibandingkan sebelum menggunakan metode pemilihan fitur. Penelitian yang dilakukan [3] menggunakan *Naïve Bayes* serta *Support Vector Machine* dan dengan menggunakan pemilihan fitur *Genetic Algorithm* yang menghasilkan akurasi dari *Naïve Bayes* sebesar 92,85% dan *Support Vector Machine* sebesar 93,03%. Pada penelitiannya di

tahun 2020, [13] menggunakan algoritma *Naïve Bayes* dan menghasilkan akurasi sebesar 73%. Pada penelitian sebelumnya hanya menggunakan 200 *dataset* dan metode yang digunakan sama, tetapi pada penelitian yang dilakukan [3] menunjukkan bahwa *Support Vector Machine* memiliki akurasi yang lebih besar dari pada *Naïve Bayes*.

Maka dari itu pada skripsi analisis sentimen ini menggunakan metode *Support Vector Machine* (SVM) yang menurut [16] merupakan salah satu metode terbaik dalam pengklasifikasian, prediksi, dan regresi. SVM memiliki akurasi yang tinggi dan bekerja sangat baik dengan dataset yang terbatas. Akurasi merupakan rasio prediksi benar positif dan negatif dengan keseluruhan data. Pada skripsi ini untuk mengetahui seberapa pengaruhnya jumlah dataset dalam proses analisis sentimen ulasan restoran dalam meningkatkan akurasi dan mencari parameter yang tepat untuk meningkatkan akurasi dari *Support Vector Machine* hingga mencapai 90%..

2. TINJAUAN PUSTAKA

2.1 Analisis Sentimen

Analisis sentimen adalah sebuah teknik atau cara yang digunakan untuk mengidentifikasi bagaimana sebuah sentimen diekspresikan menggunakan teks dan bagaimana sentimen tersebut bisa dikategorikan sebagai sentimen positif maupun sentimen negatif [10]. Pendapat yang sama juga disampaikan oleh [1], di mana analisis sentimen adalah proses yang digunakan untuk menentukan opini, emosi dan sikap yang dicerminkan melalui teks, dan biasanya diklasifikasikan menjadi opini negatif dan positif.

Dari pendapat di atas, dapat diambil kesimpulan bahwa analisis sentimen adalah sebuah proses untuk menentukan sentimen atau opini dari seseorang yang diwujudkan dalam bentuk teks dan dapat dikategorikan sebagai sentimen posisif atau negatif.

Analisis sentimen sangat diperlukan dalam menyaring opini di internet. Analisis sentimen dalam skripsi ini merupakan proses klasifikasi ulasan restoran ke dalam dua kelas, yaitu kelas sentimen positif dan negatif. Metode klasifikasi yang digunakan dalam skripsi ini adalah *Support Vector Machine*.

2.2 Restoran

Menurut arti kata dari Kamus Besar Bahasa Indonesia, restoran adalah rumah makan. Ada beberapa definisi restoran berdasarkan penelitian yang dilakukan oleh beberapa ahli, salah satu diantaranya [18] restoran adalah suatu tempat dimana pengunjung dapat menggunakan alat indera untuk menikmati pengalaman tertentu..

2.3 Permasalahan Pada Restoran

Permasalahan pada restoran salah satunya adalah kualitas layanan. Kualitas layanan menjadi isu yang dipandang sangat penting dalam memasarkan produk saat ini supaya produk dapat diterima dengan baik di pasar. Restoran harus menawarkan layanan yang mampu diterima atau dirasakan pelanggan sesuai atau melebihi apa yang diharapkan oleh pelanggan untuk menciptakan kualitas layanan yang tinggi. Menurut [5] semakin tinggi kualitas layanan yang dirasakan pelanggan dibanding harapannya, pelanggan tentu akan semakin puas. Oleh karena itu evaluasi maupun perbaikan kualitas produk atau jasa menjadi sangat penting dilakukan jika restoran ingin tetap eksis dimata pelanggannya.

2.4 *Natural Language Processing* (NLP)

Natural Language Processing (NLP) adalah suatu program komputer yang memiliki kemampuan untuk memproses bahasa manusia, baik lisan maupun tulisan yang digunakan oleh manusia dalam percakapan sehari – hari. *Natural Language Processing* (NLP) bertujuan untuk merancang dan membangun aplikasi yang memfasilitasi interaksi manusia dengan mesin dan perangkat lain melalui bahasa alami manusia [14].

2.5 *Text Mining*

Text Mining adalah satu langkah dari analisis teks yang dilakukan secara otomatis oleh komputer untuk menggali informasi yang berkualitas dari suatu rangkaian teks yang terangkum dalam sebuah dokumen [4]. Tahapan dalam *text mining* adalah *case folding, tokenizing, filtering,* dan *stemming*, tahapan ini disebut tahap *text preprocessing.* Tahap *text preprocessing* adalah tahapan dimana aplikasi melakukan seleksi data yang akan diproses pada setiap dokumen.

2.6 *Support Vectore Machine* (SVM)

Support Vector Machine (SVM) adalah suatu teknik yang relatif baru untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi. *Support Vector Machine* masuk kelas *supervised learning*, dimana dalam implementasinya perlu adanya tahap pelatihan menggunakan *sequential training* SVM dan disusul tahap pengujian [16].

Konsep klasifikasi dengan *Support Vector Machine* adalah mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua kelas data. *Support Vector Machine* mampu bekerja pada *dataset* yang berdimensi tinggi dengan menggunakan *kernel* trik. *Support Vector Machine* hanya menggunakan beberapa titik data terpilih yang berkontribusi (*support vector*) untuk membentuk model yang digunakan dalam proses klasifikasi.

Tingkat akurasi pada model yang dihasilkan oleh proses peralihan dengan SVM sangat bergantung terhadap fungsi *kernel* dan parameter yang digunakan [17]. Berdasarkan dari karakteristiknya, metode SVM dibagi menjadi dua, yaitu SVM Linier dan SVM *Non*-Linier. SVM linier merupakan data yang dipisahkan secara linier, yaitu memisahkan kedua kelas pada *hyperplane* dengan *soft margin*. Sedangkan SVM Non-Linier yaitu menerapkan fungsi dari *kernel trick* terhadap ruang yang berdimensi tinggi [15].

Persamaan *Support Vector Machine*:

$$f(x) = w.x + b \quad (1)$$

Sumber persamaan: [19]

atau

$$f(x) = \sum_{i=1}^{m} a_i y_i K(x,y) + b \quad (2)$$

Sumber persamaan: [19]

Keterangan:

w : parameter *hyperplane* yang dicari (garis yang tegak lurus antara garis *hyperplane* dan titik *support vector*)

x : titk data masukan *Support Vector Machine*

a_i : nilai bobot setiap titik data

$K(x,y)$: fungsi *kernel*

b : parameter *hyperplane* yang dicari (nilai bias)

Untuk skripsi ini menggunakan *kernel* linear dan *radial basis function* (RBF). Persamaannya:

- Linear

Menurut [11] linear *kernel* merupakan fungsi *kernel* yang paling sederhana. Linear *kernel* digunakan ketika data yang dianalisis sudah terpisah secara linear. Linear *kernel* cocok ketika terdapat banyak fitur dikarenakan pemetaan ke ruang dimensi yang lebih tinggi tidak benar – benar meningkatkan kinerja seperti pada klasifikasi teks.

$$K(x, y) = x.y \tag{3}$$

Sumber persamaan: [19]

Keterangan:

$K(x, y)$: nilai *kernel* dari data x dan data y

x : nilai data 1

y : nilai data 2

- RBF

Menurut penjelasan dari [12], RBF *kernel* merupakan fungsi *kernel* yang biasa digunakan dalam analisis ketika data tidak terpisah secara linear. RBF kernel memiliki dua parameter yaitu *Gamma* dan *Cost*. Parameter *Cost* atau biasa disebut sebagai C merupakan parameter yang bekerja sebagai pengoptimalan SVM untuk menghindari misklasifikasi di setiap sampel dalam *training dataset*. Parameter *Gamma* menentukan seberapa jauh pengaruh dari satu sampel *training dataset* dengan nilai rendah berarti "jauh", dan nilai tinggi berarti "dekat". Dengan gamma yang rendah, titik yang berada jauh dari garis pemisah yang masuk akal dipertimbangkan dalam perhitungan untuk garis pemisah. Ketika gamma tinggi berarti titik – titik berada di sekitar garis yang masuk akal akan dipertimbangkan dalam perhitungan

$$K(x, y) = exponent\left(-\frac{||x - y||^2}{2\sigma^2}\right) \tag{4}$$

Sumber persamaan: [19]

Keterangan:

$K(x, y)$: nilai *kernel* dari data x dan data y

$||x - y||$: jarak *euclidean* antara x dan y

σ : gamma

2.7 *Web Scraping*

Web scraping merupakan kegiatan yang dilakukan untuk mengambil data tertentu secara semi-terstruktur dari sebuah halaman *website*. Halaman tersebut umumnya dibangun menggunakan bahasa *markup* seperti HTML atau XML, proses akan menganalisis dokumen sebelum memulai mengambil data. Dengan melakukan *web scraping*, pengumpulan data menjadi lebih cepat, dan apabila data dikumpulkan dalam jumlah besar tidak perlu melakukan secara manual. *Web scraping* dapat dilakukan dengan menggunakan *Scrapper*, *BeautifulSoup*, *Scraper* API, dan lain – lain.

2.8 *Term Frequency – Inverse Document Frequency* (TF-IDF)

TF-IDF merupakan pembobotan yang biasa dan sering digunakan dalam pengambilan bobot dari suatu informasi dalam *text mining* [7]. Bobot TF-IDF digunakan untuk mengevaluasi seberapa pentingnya sebuah kata di dalam sebuah dokumen. *Term Frequency* (TF) yaitu semakin tinggi frekuensi kemunculan term pada sebuah dokumen maka menjadi semakin tinggi juga nilai bobot untuk term itu sendiri. IDF merupakan kebalikan dari TF, semakin tinggi frekuensi kemunculan term maka nilai bobot term itu sendiri menjadi semakin kecil.

Dalam skripsi ini menggunakan TF-IDF yang sudah disediakan oleh *library* scikit-learn yaitu TfIdfVectorizer. Salah satu parameter yang ada dalam TfIdfVectorizer adalah n-gram. N-gram merupakan potongan n-karakter yang diperoleh dari sebuah kalimat. Pada skripsi ini, n-gram digunakan dengan pemecahan berdasarkan per-kata. Pada umumnya, n-gram yang utuh didapat dengan menambahkan blank di awal dan di akhir suatu kata. Misalnya, terdaapt kata "TEKS" yang dapat diuraikan ke dalam beberapa n-gram berikut yang dimana "_" mempresentasikan blank [6].

Unigram : T, E, K, S

Bigram : _T, TE, EK, KS, dan S_

Trigram : _TE, TEK, EKS, KS_, dan S__

Persamaan TF-IDF yang digunakan pada skripsi ini dapat dilihat pada Persamaan 5.

$$w_{i,j} = tf_{i,j} \log\left({}^{N}/_{df_i}\right) \tag{5}$$

keterangan:

$w_{i,j}$ = bobot dokumen ke-i terhadap kata ke-j

$tf_{i,j}$ = banyaknya kemunculan kata i yang dicari pada sebuah dokumen j

N = jumlah semua dokumen yang ada

df_i = banyaknya dokumen yang mengandung kata ke-i

2.9 *Confusion Matrix &* F1-*Score*

Confusion Matrix adalah sebuah metode yang biasa digunakan untuk perhitungan akurasi [8]. Keakuratan hasil dievaluasi dengan nilai *recall*, *precision*, *accuracy*, dan *error rate*. Dimana *recall* (*True Positive Rate*) merupakan rasio identifikasi benar positif dibandingkan keseluruhan data yang benar positif. *Precision* (*Positive Predictive Value*) merupakan rasio identifikasi benar positif dibandingkan keseluruhan hasil yang diidentifikasi positif. *Accuracy* merupakan rasio identifikasi benar (positif dan negatif) dengan keseluruhan data. *Error rate* merupakan rasio identifikasi salah dengan keseluruhan data. F1-*score* merupakan rata – rata dari *precision* dan *recall*.

Rumus – rumus untuk mendapatkan nilai *recall*, *precision*, *accuracy*, *error rate*, f1-score sebagai berikut:

$$recall = \frac{TP}{TP+FN} \tag{2}$$

$$precision = \frac{TP}{TP+FP} \tag{3}$$

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$error\ rate = \frac{FP+FN}{TP+TN+FP+FN} \quad (5)$$

$$F1\ score = 2 * \frac{precision*recall}{precision+recall} \quad (6)$$

2.10 Tinjauan Studi

Berikut tinjauan studi dari beberapa penelitian yang sudah pernah dilakukan sebelumnya:

1. Analisis Sentimen Pada *Review* Restoran Dengan Teks Bahasa Indonesia Mengunakan Algoritma Naive Bayes [9]

• Masalah yang diangkat pada penelitian ini adalah berbagai konten web meliputi opini subjektif serta informasi yang objektif, membuat orang - orang mengumpulkan informasi tentang produk dan jasa yang mereka ingin beli. Karena banyaknya informasi yang ada dalam bentuk teks tanpa ada skala numerik membuat suatu informasi sulit dievaluasi secara efisien tanpa membaca secara lengkap.

• Hasil dari penelitian tersebut adalah pembuatan sistem yang membantu para calon pembeli dalam mengambil keputusan saat ingin mencari restoran dengan menggunakan algoritma *Naive Bayes* dengan metode pemilihan fitur *Genetic Algorithm*.

• Perbedaan penelitian dari pembuatan skripsi ini adalah metode dan jumlah data yang digunakan untuk melakukan klasifikasi.

2. Analisis Sentimen Menggunakan Algoritma *Naive Bayes* Terhadap *Review* Restoran di Singapura [13]

• Masalah yang diangkat pada penelitian ini adalah banyaknya variasi restoran menjadi suatu masalah bagi pengunjung dalam memilih restoran yang ingin dikunjungi, sehingga pengunjung melihat rekomendasi atau penilaian pengunjung lain.

• Hasil dari penelitian tersebut adalah pembuatan sistem yang dapat menentukan klasifikasi dari suatu *review* ke dalam dua kategori yaitu positif dan negatif.

• Perbedaan penelitian dari pembuatan skripsi ini adalah metode dan jumlah data yang digunakan untuk melakukan klasifikasi.

3. Komparasi Algoritma *Support Vector Machine* Dan *Naive Bayes* Dengan Algoritma Genetika Pada Analisis Sentimen Calon Gubernur Jabar 2018-2023 [3]

• Masalah yang diangkat pada penelitian ini adalah banyaknya masyarakan yang memberikan cuitan di Twitter pada masa kampanye calon gubernur Jawa Barat 2018 – 2023 untuk membrikan dukungan atau tidak. Membaca keseluruhan cuitan yang tersebar akan memakan waktu dan membingungkan dalam pengambilan keputusan.

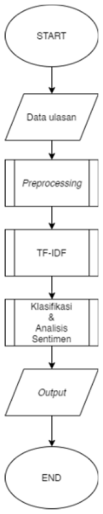
• Hasil dari penelitian tersebut adalah pembuatan sistem yang dapat melakukan klasifikasi teks dalam pola negatif atau positif dari cuitan mengenai calon gubernur Jawa Barat 2018- 2023.

• Perbedaan penelitian dari pembuatan skripsi ini adalah metode yang digunakan dan data yang digunakan dalam melakukan klasifikasi.

3. METODE PENELITIAN

3.1 Analisis Sistem

Analisa sistem membahas mengenai alur proses dari program meliputi bagaimana bentuk input data hingga dapat menghasilkan *output* yang diharapkan. Secara garis besar sistem dimulai dari *input* yang berasal dari pengumpulan data dari hasil *web scraping* pada berbagai restoran di Tripadvisor, kemudian dilakukan *preprocessing* data dan pembuatan model SVM. *Training* dan *testing* sentimen hingga menghasilkan output berdasarkan model yang telah dibuat. Proses alur sistem ditunjukkan pada Gambar 1.



Gambar 1. Proses alur sistem aplikasi

3.2 Pengumpulan Data

Data yang digunakan dalam skripsi ini adalah kumpulan ulasan berbahasa Indonesia dari berbagi restoran. Data ulasan didapatkan dari proses *web scraping* menggunakan *BeautifulSoup* pada *website* Tripadvisor. Jumlah total data adalah 300 ulasan berbahasa Indonesia, dengan 150 ulasan positif dan 150 ulasan negatif. Data yang didapatkan disimpan pada file dengan ekstensi .csv.

3.3 Pembagian Data *Training* dan *Testing*

Dataset yang tersedia dipecah menjadi data *training* dan data *testing* dengan jumlah perbandingan 4:1 yaitu sebanyak 80% sebagai data *training* dan sebanyak 20% sebagai data *testing*.

3.4 *Preprocessing*

Proses *preprocessing* dilakukan untuk mengelolah data yang bervariasi karena data ini dapat mempengaruhi proses *training* dan *testing*. Proses *preprocessing* memanfaatkan *library Natural Language Toolkit* (NLTK). Proses pertama dalam *training* adalah *case folding* yang bertujuan untuk menjadikan semua teks menjadi huruf kecil. Proses kedua adalah tokenisasi yang bertujuan untuk memotong kalimat dan menghilangkan karakter seperti petik tunggal (‘), titik (.), titik dua (:) atau lainnya. Proses ketiga adalah *stopword removal* yang bertujuan untuk menghapus kata – kata umum yang tidak memiliki arti atau makna. Proses terkahir adalah *stemming* yang bertujuan untuk menjadikan kata – kata yang sudah diproses sebelumnya menjadi kata dasar.

3.5 TF-IDF

Proses TF-IDF merupakan proses pemberian bobot pada dokumen dan mengubah data yang awalnya adalah suatu atribut teks menjadi *matrix* yang berisi numerik.

3.6 *Training & Testing*

Proses *training* dilakukan untuk mempelajari data teks agar ketika mendapat teks baru sistem dapat memprediksi teks baru sesuai dengan teks yang sudah dipelajarinya. Proses ini melakukan *split* pada *dataset* menjadi data *training* dan data *testing*. Data *testing* digunakan untuk memberikan tes prediksi pada sistem untuk menghasilkan akurasi model.

4. PENGUJIAN

4.1 Pengujian Jumlah *Dataset* & *Preprocessing*

Pada tahap ini dilakukan pengujian terhadap jumlah *dataset* dan *preprocessing*. Pengujian dilakukan dengan membandingkan nilai testing dengan 50, 100, 200, dan 300 data ulasan dari berbagai restoran serta sebelum dan sesudah *preprocessing*. Perbandingan dilakukan dengan membandingkan nilai akurasi dan F1-*Score* saja. Pengujian ini dilakukan pada data ulasan berbahasa Indonesia dari berbagai restoran.

Pengujian ini dilakukan dengan melakukan proses *testing* pada data dengan SVM *classifier* dengan dua *kernel* yaitu rbf dan linear. Hasil berupa nilai akurasi dan f1-*score* sebelum dan sesudah dilakukan *preprocessing* dapat dilihat pada Tabel 1 dan Tabel 2.

Tabel 1. Nilai Akurasi Sebelum dan Sesudah *Preprocessing* pada 50, 100, 200, dan 300 *Dataset*

Kernel	*Dataset*	Akurasi Sebelum *Preprocessing*	Akurasi Sesudah *Preprocessing*
RBF	50	0.40	0.45
	100	0.45	0.55
	200	0.82	0.83
	300	0.85	0.90
Linear	50	0.70	0.70
	100	0.65	0.75
	200	0.84	0.84
	300	0.87	0.91

Tabel 2. Nilai F1-*Score* Sebelum dan Sesudah *Preprocessing* pada 50, 100, 200, dan 300 *Dataset*

Kernel	*Dataset*	F1-score Sebelum *Preprocessing*	F1-score Sesudah *Preprocessing*
RBF	50	0.40	0.45
	100	0.45	0.55
	200	0.82	0.83
	300	0.85	0.90
Linear	50	0.70	0.70
	100	0.65	0.75
	200	0.84	0.84
	300	0.87	0.91

4.2 Pengujian Parameter TfidfVectorizer

Pada tahap ini dilakukan pengujian terhadap parameter TfidfVectozier yang dikombinasikan satu per satu untuk menemukan parameter TF-IDF terbaik. Pengujian dilakukan bersamaan dengan *cross validation* sebanyak 5 lipatan. Pengujian ini dilakukan terhadap jumlah dataset dan dua kernel SVM yang digunakan untuk pengujian sebelumnya. Penggunaan *preprocessing* juga digunakan kembali.

Pengujian ini mencoba parameter TF-IDF dengan kernel linear dan rbf pada 50, 100, 200, dan 300 data ulasan restoran yang sudah melakukan proses *preprocessing*. Hasil dari pengujian ini dapat dilhat pada Tabel 3.

Tabel 3. Parameter TF-IDF Linear pada 50,100, 200, dan 300 Data Ulasan Restoran

Jumlah *Dataset*	Max_df	Min_df	N-gram	Norm	Nilai
50	0.75	0.05	(1, 1)	l2	0.760
100	0.75	0.05	(1, 1)	l2	0.800
200	0.75	0.05	(1, 2)	l2	0.840
300	0.75	0.05	(1, 2)	l2	0.860

4.3 Pengujian Parameter SVM

Pada tahap pengujian ini dilakukan *training* data pada *kernel* linear dan rbf. *Kernel* linear memiliki parameter C, yang dimana parameter C digunakan untuk menentukan kerapatan margin antar *support vector*, semakin besar nilai C maka kerapatan margin juga semakin dekat, tetapi jika nilai C terlalu besar dapat menyebabkan *overfit*. *Kernel* rbf memiliki parameter gamma, yang dimana parameter gamma digunakan untuk menentukan seberapa jauh pengaruh yang dicapai oleh satu contoh pelatihan.

Pengujian menggunakan parameter C dengan nilai 0.1, 1, 10, dan 100 untu *kernel* linear, sedangkan untuk *kernel* rbf pengujian parameter dengan gamma 0.1, 1, dan 10. Hasil dari pengujian dapat dilihat pada Tabel 4.

Tabel 4. Nilai Akurasi dari 50, 100, 200, dan 300 Dengan *Kernel* Linear dan RBF

Jumlah *Dataset*	*Kernel*	C	*Gamma*	Nilai
50	RBF	1.0	0.01	0.833
100	Linear	1.0	-	0.853
200	Linear	1.0	-	0.860
300	Linear	1.0	-	0.875

4.4 Pengujian Model SVM dan TF-IDF Terhadap Akurasi dan F1-Score

Pada pengujian ini dilakukan dengan menggunakan parameter yang sudah didapat sebelumnya. Dari parameter yang sudah didapat dilakukan pengujian kembali untuk medapatkan parameter yang terbaik. Hasil dari pengujian dapat dilihat pada Tabel 5.

Tabel 5. Nilai Parameter SVM dan TF-IDF

Kernel	C	Min-Df	Max-Df	Ngram	Norm	Nilai
Linear	1	0.05	0.75	(1, 1)	l2	0.823
Linear	1	0.05	0.75	(1, 1)	l2	0.830
Linear	1	0.05	0.75	(1, 2)	l2	0.850
Linear	1	0.05	0.75	(1, 2)	l2	0.880

Dari paramater yang sudah didapat, selanjutnya dilakukan pengujian pada SVM dan TF-IDF terhadap akurasi dan f1-score. Hasil dari pengujian dapat dilihat pada Tabel 6.

Tabel 6 Hasil dari pengujian SVM & TF-IDF terhadap f1-score & akurasi

Jumlah Dataset	F1-Score	Akurasi
50	0.70	0.70
100	0.70	0.70
200	0.92	0.92
300	0.93	0.93

Berdasarkan pengujian yang telah dilakukan dengan menggunakan parameter yang paling baik menunjukkan bahwa ada peningkatan pada nilai akurasi dan f1-*score*, dengan akurasi sebesar 0.93 dan f1-*score* 0.93.

4.5 Pengujian Riil

Pengujian riil adalah pengujian yang dilakukan dengan membandingkan hasil dari *Support Vector Machine* dengan opini penulis dari data yang diuji. Pengujian yang pertama menggunakan data ulasan dari Fajar Chinese *Restaurant* – Tunjungan Plaza. Pada Gambar 2 dapat dilihat hasil pengujian menggunakan *Support Vector Machine* dan Tabel 7 adalah perbedaan hasil dari *Support Vector Machine* dengan opini penulis dari data ulasan Fajar Chinese Restaurant – Tunjungan Plaza.

Gambar 2. Hasil pengujian dengan menggunakan *Support Vector Machine* pada Fajar Chinese *Restaurant* – Tunjungan Plaza

Tabel 7. Perbedaan hasil dari *Support Vector Machine* dengan opini penulis pada data ulasan Fajar Chinese *Restaurant* – Tunjungan Plaza

Ulasan	SVM	Penulis
Chinese food fajar restoran ini sudah terkenal enaknya dari sejak saya masih kecil 20 lebih tahun yang lalu. Semua masakannya rata-rata enak enak semua, mulai koloke, burung dara, bakmie. Semuanya saya suka banget. sudah menjadi favorit saya sejak masih anak-anak	Negatif	Positif
Telah di sini selama tahun baru Cina. Tempat yang sangat lengkaptua restoran China terkenal di sby. Tapi. Mereka berhasil untuk menjaga kualitasSemua makanan yang sangat baikHarga. Sangat layak	Negatif	Positif

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan pengujian yang telah dilakukan pada data ulasan dari berbagai restoran dengan Bahasa Indonesia, diambil kesimpulan sebagai berikut:

1. Penggunaan *preprocessing* dapat meningkatkan hasil akurasi dan f1-*score*.
2. Menambah jumlah *dataset* yang lebih banyak dapat menghasilkan akurasi dan f1-*score* yang lebih baik.
3. Parameter SVM terbaik adalah *kernel* linear dan parameter C sebesar 1.
4. Parameter TF-IDF terbaik adalah min_df sebesar 0.05, max_df sebesar 0.75, norm l2, n-gram (1, 2).
5. Akurasi tertinggi yang didapat dari model SVM sebesar 0.93.
6. *Precision* tertinggi yang didapat dari model SVM sebesar 0.93.
7. *Recall* tertinggi yang didapat dari model SVM sebesar 0.93.
8. *Error rate* terendah yang didapat dari model SVM sebesar 0.06.
9. F1-Score tertinggi yang didapat dari model SVM sebesar 0.93.

5.2 Saran

Dari penelitian yang dilakukan ada beberapa saran yang mungkin berguna untuk penelitian selanjutnya. Berikut saran yang ada:

1. Jumlah *dataset* dapat ditambahkan agar dapat menghasilkan hasil yang lebih baik lagi.
2. Dapat menganalisa ulasan restoran dengan bahasa yang lain.
3. Aplikasi mungkin bisa disempurnakan sehingga dapat diakses melalui *website* dan *mobile*

6. DAFTAR REFERENSI

[1] Coletta, L. F., da Silva, N. F., Hruschka, E. R., & Hruschka, E. R. 2014. Combining Classification and Clustering for Tweet Sentiment Analysis. *Brazilian Conference on Intelligent Systems*, 210 - 215.

[2] Govindarajan, V., Anthony, R., Hartmann, F., Kraus, K., & Nilsson, G. 2013. *EBOOK: Management Control Systems: European Edition.* McGraw Hill.

[3] Gunawan, D., Riana, D., Ardiansyah, D., Akbar, F., & Alfarizi, S. 2020. Komparasi Algoritma Support Vector Machine Dan Naive Bayes Dengan Algoritma Genetika Pada Analisis Sentimen Calon Gubernur Jabar 2018-2023. *Jurnal Teknik Komputer AMIK BSI, VI*(1), 121-129.

[4] Han, J., & Kamber, M. 2006. Classification and prediction. *Data mining: Concepts and techniques*, 347-350.

[5] Istijanto, M. M. 2005. *Aplikasi Praktis Riset Pemasaran.* Gramedia Pustaka Utama.

[6] Lieaharyani, D. C. 2015. Automatic Essay Scoring System Using N-Gram and Cosine Similarity for Gamification Based E-Learning.

[7] Maulina, D., & Sagara, R. 2018. Klasifikasi artikel hoax menggunakan support vector machine linear dengan pembobotan term frequency-Inverse document frequency. *Jurnal Mantik Penusa, 2*(1).

[8] Melita, R., Amrizal, V., Suseno, H. B., & Dirjam, T. 2018. Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Syarah Umdatil Ahkam). *Jurnal Teknik Informatika, 11*(2), 149-164.

[9] Muthia, D. A. 2017. Analisis Sentimen Pada Review Restoran Dengan Teks Bahasa Indonesia Mengunakan Algoritma Naive Bayes. *Jurnal Ilmu Pengetahuan dan Teknologi Komputer*, 39-45.

[10] Nasukawa, T., & Yi, J. 2003. Sentiment Analysis: Capturing Favorability Using Natural Language Processing. *In Proceedings of the 2nd International Conference on Knowledge Capture*, 70 - 77.

[11] Ningrum, H. C. 2018. Perbandingan Metode Support Vector Machine (SVM) Linear, Radial Basis Function (RBF), dan Polinomial Kernel Dalam Klasifikasi Bidang Studi Lanjut Piliihan Alumni UII.

[12] Patel, S. 2018. *Chapter 2 : SVM (Support Vector Machine)—Theory*. Retrieved from Medium: https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

[13] Permadi, V. A. 2002. Analisis Sentimen Menggunakan Algoritma Naive Bayes Terhadap Review Restoran di Singapura. *Jurnal Buana Informatika, 11*(2), 141-151.

[14] Pustejovsky, J., & Stubbs, A. 2012. *Natural Language Annotaion for Machine Learning: A guide to corpus-building for application.* O'Reilly Media, Inc.

[15] Rachman, F., & S. W., P. 2012. Perbandingan Klasifikasi Tingkat Keganasan Breast Cancer Dengan Menggunakan Regresi Logistik Ordinal Dan Support Vector Machine (SVM).

[16] Santoso, B. 2007. *Data Mining Teknik Pemanfaatan Untuk Keperluan Bisnis.* Yogyakarta: Graha Ilmu.

[17] Siagian, R. Y. 2011. Klasifikasi Parket Kayu Jati Menggunakan Metode Support Vector Machine (SVM).

[18] Walker, J. R., & Lundberg, D. E. 2005. *The restaurant: from concept to operation* (4th ed.). Hoboken: Wiley.

[19] Wang, H., & Hu, D. 2005. Comparison of SVM and LS-SVM for Regression. *International Conference on Neural Network and Brain*, 279 - 283.

[20] Zhang, Z., Ye, Q., Zhang, Z., & Li, Y. 2011. Sentiment classification of Internet restaurant reviews written in Cantonese. *Expert Systems with Applications, 38*(6), 7674-7682.