Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik20185867
Vol.5, No.5 Oktober 2018, hlm.537-548 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 51/E/KPT/2017 e-ISSN:2528-6579

SUPPORT VECTOR MACHINE BERBASIS *FEATURE SELECTION* UNTUK *SENTIMENT ANALYSIS* KEPUASAN PELANGGAN TERHADAP PELAYANAN WARUNG DAN RESTORAN KULINER KOTA TEGAL

Oman Somantri[1], Dyah Apriliani[2]

[1,2]Jurusan Teknik Informatika, Politeknik Harapan Bersama Tegal, Indonesia
Email: [1]oman.somantri@poltektegal.ac.id, [2]dyah.apriliani90@gmail.com

(Naskah masuk: 16 Juni 2018, diterima untuk diterbitkan: 17 Oktober 2018)

Abstrak

Setiap pelanggan pasti menginginkan sebuah pendukung keputusan dalam menentukan pilihan ketika akan mengunjungi sebuah tempat makan atau kuliner yang sesuai dengan keinginan salah satu contohnya yaitu di Kota Tegal. *Sentiment analysis* digunakan untuk memberikan sebuah solusi terkait dengan permasalahan tersebut, dengan menereapkan model algoritma S*upport Vector Machine* (SVM). Tujuan dari penelitian ini adalah mengoptimalisasi model yang dihasilkan dengan diterapkannya *feature selection* menggunakan algoritma *Information Gain* (IG) dan *Chi Square* pada hasil model terbaik yang dihasilkan oleh SVM pada klasifikasi tingkat kepuasan pelanggan terhadap warung dan restoran kuliner di Kota Tegal sehingga terjadi peningkatan akurasi dari model yang dihasilkan. Hasil penelitian menunjukan bahwa tingkat akurasi terbaik dihasilkan oleh model SVM-IG dengan tingkat akurasi terbaik sebesar 72,45% mengalami peningkatan sekitar 3,08% yang awalnya hanya 69.36%. Selisih rata-rata yang dihasilkan setelah dilakukannya optimasi SVM dengan *feature selection* adalah 2,51% kenaikan tingkat akurasinya. Berdasarkan hasil penelitian bahwa *feature selection* dengan menggunakan *Information Gain (IG)* (SVM-IG) memiliki tingkat akurasi lebih baik apabila dibandingkan SVM dan *Chi Squared* (SVM-CS) sehingga dengan demikian model yang diusulkan dapat meningkatkan tingkat akurasi yang dihasilkan oleh SVM menjadi lebih baik.

Kata kunci: *Sentiment Analysis, Support Vector Machine (SVM), feature selection, Information Gain (IG), Chi Square*

SUPPORT VECTOR MACHINE BASED ON FEATURE SELECTION FOR SENTIMENT ANALYSIS CUSTOMER SATISFACTION ON CULINARY RESTAURANT AT TEGAL CITY

Abstract

The Customer needs to get a decision support in determining a choice when they're visit a culinary restaurant accordance to their wishes especially at Tegal City. Sentiment analysis is used to provide a solution related to this problem by applying the Support Vector Machine (SVM) algorithm model. The purpose of this research is to optimize the generated model by applying feature selection using Informatioan Gain (IG) and Chi Square algorithm on the best model produced by SVM on the classification of customer satisfaction level based on culinary restaurants at Tegal City so that there is an increasing accuracy from the model. The results showed that the best accuracy level produced by the SVM-IG model with the best accuracy of 72.45% experienced an increase of about 3.08% which was initially 69.36%. The difference average produced after SVM optimization with feature selection is 2.51% increase in accuracy. Based on the results of the research, the feature selection using Information Gain (SVM-IG) has a better accuracy rate than SVM and Chi Squared (SVM-CS) so that the proposed model can improve the accuracy of SVM better.

Keywords: *Sentiment Analysis, Support Vector Machine (SVM), feature selection, Information Gain (IG), Chi Square*

1. PENDAHULUAN

Salah satu ciri khas yang dimiliki oleh Indonesia adalah keberagaman akan makananan kuliner yang dimilikinya, dimana hampir setiap daerah memiliki ciri khas makanan kuliner yang berbeda-beda dan hal ini tentunya menjadikan banyak para penikmat makanan kuliner yang berasal dari negara-negara lain maupun Indonesia sendiri untuk mencari referensi terkait dengan tempat terbaik yang direkomendasikan oleh orang-orang atau pelanggan yang pernah singgah ditempat tersebut yang menjajakan makanan kuliner ciri khas pada setiap daerah di Indonesia. Tempat-tempat yang direkomendasikan adalah berbagai macam restoran serta warung-warung kuliner yang menjual makanan ciri khas daerah tersebut, salah satunya adalah daerah Kota Tegal. Kota ini merupakan sebuah daerah yang berada dikawasan wilayah pantura provinsi Jawa Tengah yang memiliki banyak sekali makanan kuliner yang menjadi ciri khas kota ini.

Perkembangan teknologi saat ini memungkinkan orang untuk dapat dengan mudah mendapatkan sebuah informasi, baik itu dari media sosial, website maupun yang lainnya. Salah satu informasi yang bisa didapatkan terkait dengan rekomendasi warung dan restoran kuliner terbaik di Indonesia disetiap daerah termasuk di Kota Tegal adalah dengan membaca komentar-komentar yang ditulis oleh orang-orang yang terindikasi pernah singgah ditempat tersebut pada halaman *website* maupun media sosial. Melalui komentar-komentar dan pendapat orang-orang yang pernah merasakan makanan kuliner ditempat yang pernah disinggahi maka dapat dijadikan sebagai pendukung keputusan para pelanggan yang dalam hal ini penikmat kuliner untuk datang ketempat tersebut serta dijadikan pula sebagai pendukung keputusan para pemilik warung dan restoran kuliner untuk dijadikan sebagai acuan tingkat keberhasilan bentuk pelayanan terhadap pelanggannya (Reyes and Rosso, 2012).

Permasalahan yang terjadi adalah terkadang para pelanggan tidak mungkin untuk dapat membaca komentar-komentar yang terlalu banyak untuk mendapatkan sebuah keputusan rekomendasi pilihan tempat mana yang terbaik, selain itu permasalahanpun terjadi dari pihak pemilik warung dan restoran kuliner yang ingin mandapatkan sebuah data terkait dengan komentar-komentar para penikmat kuliner terhadap tempatnya untuk dapat menentukan sebuah keputusan terkait dengan pelayanan yang diberikan sesuai dengan keinginan pelanggan atau masih perlu adanya sebuah peningkatan pelayanan baik itu dari segi rasa makanan, kenyamanan tempat, maupun pelayanan ditempat tersebut. Terkait dengan permasalahan yang ada, diperlukan sebuah metode yang dapat membantu untuk menganalisis terkait dengan komentar-komentar tersebut. Solusi yang dilakukan adalah diterapkannya model *sentiment analysis* (SA).

Metode s*entiment analysis* (SA) adalah sebuah proses dalam memahami, meng-ekstrak, dan mengolah sebuah data yang berupa tekstual yang didapatkan dari kalimat opini-opini yang berkerja secara otomatis sehingga didapatkan informasi sentiment yang terkandung didalamnya. *Sentiment analysis* dilakukan sebagai upaya untuk melihat sebuah pendapat atau opini terhadap sebuah objek atau permasalahan oleh seseorang yang akhirnya nanti mempunyai nilai kecenderungan apakah menilai positif atau negatif. *Sentiment analysis* mengacu pada aplikasi dari *Natural Language Processing* (NLP), *Computational Linguistic* dan *text analytics* untuk mengidentifikasi dan mengekstrak informasi subjektif dari materil sumber (Liu, 2010). Sesuai dengan kelebihan yang dimiliki oleh SA, berbagai macam penelitian dan *review* telah dilakukan mengenai penggunaan SA ini untuk mendapatkan sebuah pendukung keputusan diantaranya mengarah pada hal *subjectivity classification*, klasifikasi sentiment, deteksi opini spam, dan review mengukur kegunaan, *aspect extraction* serta *lexicon creation* (Ravi and Ravi, 2015).

Dalam penerapan *Sentiment Analysis* terdapat beberapa metode *machine learning* yang sering digunakan, diantaranya adalah *Decision Tree classifier, Neural Network (NN), dan Naïve bayes (NB) serta Bayesian Network, selain itu Support Vector Machines (SVM)* dan *Maximum Entropy* (Medhat, Hassan and Korashy, 2014). Diantara beberapa teknik metode yang digunakan, SVM adalah salah satu metode terbaik yang sering digunakan karena tingkat akurasinya menghasilkan lebih baik (Tripathy, Agrawal and Rath, 2015). Metode SVM memiliki beberapa kelebihan salah satunya adalah dapat diterapkan pada data yang berdimensi tinggi, disisi lain selain kelebihannya kekurangan yang dimiliki SVM adalah masih sulit digunakan untuk data yang jumlah besar.

Permasalahan yang terjadi pada sebuah klasifikasi *sentiment analysis* berbasiskan text adalah begitu banyaknya atribut yang digunakan pada dataset yang digunakan. Umumnya *atribute* dari sentiment teks sangatlah besar sehingga apabila seluruh atribut yang ada digunakan maka hal tersebut akan mengurangi kinerja dari *classifier* sehingga akan menjadikan tingkat akurasi yang dihasilkan menjadi rendah (Wang dkk., 2013). Untuk mengatasi permasalahan tersebut maka perlu sebuah cara yang dapat mengoptimalkan kinerja sistem *classifier* yang dibuat yaitu dengan *feature selection* seperti salah satunya dilakukan oleh (Somantri and Khambali, 2017) pada *text mining*, serta klasifikasi dokumen teks seperti yang dilakukan oleh (Aminudin, SN and Ahmad, 2018; Wijoyo dkk., 2017). *Feature selection* bekerja berdasarkan proses pengurangan ruang-ruang fitur yang tidak relevan dengan cara mengeliminir setiap *atribute* yang tidak relevan tersebut (Koncz and Paralic, 2011). *Information Gain* (IG) dan *Chi-*

Square merupakan salah satu metode algoritma digunakan untuk *feature selection*. Kedua algoritma tersebut merupakan algoritma *feature selection* yang mempunyai kemampuan meningkatkan hasil lebih baik apabila dibandingkan dengan metode lainnya (Tan and Zhang, 2008).

Terdapat beberapa penelitian sebelumnya terkait dengan analisis opini yang dilakukan oleh para peneliti. Salah satunya adalah (Kang, Yoo and Han, 2012) melakukan penelitian terkait dengan *sentiment analysis* yang digunakan untuk *review* restoran dengan menggunakan *Senti-lexicon* dan algoritma *Naïve bayes* (NB) yang telah di-*improved*. Pada penelitian ini diusulkan sebuah *improved* algoritma *Naïve Bayes* sebagai metode yang digunakan kemudian hasilnya dikomparasi dengan SVM, dari hasil penelitian menyimpulkan bahwa NB-*Improved* menghasilkan tingkat presisi yang lebih baik.

Penelitian dilakukan oleh (Zhang dkk., 2011) dengan melakukan proses review terhadap restoran yang bertuliskan nama Canton atau mandarin di internet, pada penelitian ini komparasi nilai hasil yang dihasilkan oleh *Support Vector Machine* (SMV) dan *Naive Bayes* (NB) untuk melakukan *sentiment classification*. Hasil penelitian menunjukan bahwa NB lebih baik dibandingkan dengan SVM. Penelitian selanjutnya dilakukan oleh (Di Caro and Grella, 2013), dalam penelitian ini *sentiment analysis* dilakukan dengan melalui *dependency parsing*. Hasil eksperimen yang telah dilakukan kemudian hasilnya dievaluasi menggunakan sebuah *dataset review* restoran. Selanjutnya penelitian dilakukan oleh (Robaldo and Di Caro, 2013), melakukan penelitian terkait dengan *opinion mining*. Pada penelitian ini mengusulkan *OpnionMining –ML*, yaitu sebuah formulasi berbasis XML baru untuk menandai ekspresi tekstual yang menyampaikan pendapat tentang objek yang dianggap relevan. Penelitian dilakukan dengan menggunakan data yang berasal dari website restoran makanan spageti yang terdapat *review* didalamnya.

Pada penelitian yang akan dilakukan ini sedikit berbeda dengan apa yang sudah dilakukan oleh beberapa peneliti sebelumnya, pada penelitian ini proses analisis *sentiment* dilakukan untuk menganalisis tingkat kepuasan pelanggan terhadap pelayanan warung dan restoran kuliner di Kota Tegal dengan cara mengklasifikasikannya antara yang beropini positif dan negatif dengan menerapkan *Support Vector Machines (SVM)* sebagai model yang digunakan. Pada penelitian ini istilah positif dan negatif diganti menjadi dua kategori klasifikasi yaitu "rata-rata" dan "bagus". Berdasarkan kelemahan yang terdapat pada algoritma tersebut, maka untuk dapat meningkatkan tingkat akurasi yang dihasilkan dilakukan optimasi dengan menggunakan *feature selection.* Algoritma yang diusulkan adalah dengan menggunakan *Information Gain* (IG) dan *chi-square*, sehingga diharapkan adanya peningkatan tingkat akurasi.

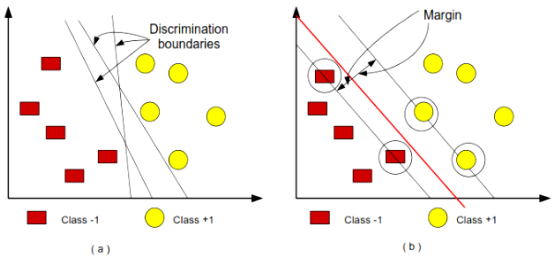
Tujuan dari penelitian ini adalah diperolehnya sebuah model terbaik untuk *sentiment analysis* dari diterapkannya metode optimasi *feature selection* yang diterapkan pada *Support Vector Machine* (SVM) digunakan untuk mengklasifikasi sentiment tingkat kepuasan para pelanggan terhadap warung dan restoran kuliner Kota Tegal sehingga terjadinya peningkatan akurasi dari model yang dihasilkan.

2. SUPPORT VEKTOR MACHINE & FEATURE SELECTION

2.1. Support Vector Machine (SVM)

Sebagai salah satu algoritma klasifikasi yang sering digunakan, *Support Vector Machine* (SVM) bekerja dengan cara mencari sebuah *hyperline* atau garis pembatas pemisah antar kelas yang mempunyai margin atau jarak antar *hyperlane* dengan data paling terdekat pada setiap kelas yang paling besar. Algoritma SVM sebenarnya dasarnya digunakan untuk proses klasifikasi antara dua kelas atau *binary classification*, sesuai dengan perkembangannya SVM digunakan untuk klasifikasi *multi-class* yaitu dengan cara kombinasi antara beberapa binary classifier (Jiu-Zhen Liang, 2004).

Sejak pertama kali dikembangkan oleh Boser, Guyon & Vapnik pada tahum 1992, SVM saat ini banyak yang menggunakannya. Konsep SVM pada dasarnya adalah upaya pencarian nilai *hyperline* yang terbaik pemisah antara dua buah *class* dalam *input space*, gambaran proses pemisahan tersebut seperti digambarkan pada Gambar 1.



Gambar 1. Hyperlane Class -1 dan +1

Pada Gambar 1 memperlihatkan 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.

2.2. Feature Selection

Feature selection merupakan sebuah cara untuk dapat menjadikan sebuah pengklasifikasi lebih efektif dan efisien serta lebih baik dengan cara mengurangi jumlah data-data yang dianalisis, atau dengan mengidentifikasi fitur-fitur yang sesuai sebagai bahan

pertimbangan pada proses pembelajaran (Moraes, Valiati and Gavião Neto, 2013). Pada *feature selection* ini terdapat dua jenis tipe utama dala melakukan proses seleksi fitur pada *machine learning*, diantaranya adalah metode *wrapper* dengan menggunakan beberapa algoritma sebagai pengukuran akurasi klasifikasi, dan metode *filter* (Chen dkk., 2009).

a. Information Gain (IG)

IG merupakan salah satu algoritma terbaik yang digunakan sebagai *feature selection*. Untuk menghitung *Information Gain* dihitung dengan menggunakan persamaan:

$$Info\ (D) = -\sum_{i=1}^{c} \quad p_i log_2(p_i) \qquad (1)$$

dimana:
c: jumlah nilai yang ada pada atribut target (jumlah kelas klasifikasi)
pi: jumlah *sample* untuk kelas *i*

$$Info_A(D) = \sum_{j=1}^{V} \left(\frac{Dj}{D}\right) x\ Info\ (D_j) \qquad (2)$$

Untuk mengukur efektifitas suatu atribut dalam pengklasifikasin data dapat dihitung dengan persamaan:

$$Gain(A) = |\ Info(D) - Info_A(D)\ | \qquad (3)$$

b. *chi-square*

Chi-square adalah satu metode yang masuk kedalam tipe dari seleksi fitur *supervised*, dimana mampu menghilangkan fitur-fitur dengan tanpa mengurangi dari tingkat akurasi yang dihasilkan. Untuk mengukur sebuah nilai *dependence* dari dua variable digunakan persamaan (4).

$$X^2(\ t,c\ ) = \frac{N\ x\ (\ AD-CB)2}{(\ A+C)\ x\ (B+D) x\ (\ A+B)\ x\ (\ C+D)} \qquad (4)$$

dimana:
A: jumlah kali fitur t dan ketegori *c* terjadi,
B: jumlah kali *t* terjadi tanpa *c*,
C: jumlah kali *c* terjadi tanpa *t*,
C: jumlah berapa kali *c* terjadi tanpa *c*,
D: jumlah kali tidak *c* atau *c* terjadi,
N: jumlah kasus.

3. METODE PENELITIAN

3.1. Dataset

Dataset pada penelitian ini menggunakan data yang diambil dari situs www.tripadvisor.co.id berupa data teks yang berisi komentar-komentar pengunjung web terhadap warung dan restoran kuliner di kota Tegal antara tahun 2017 s.d 2018 (TripAdvisor LLC, 2017). Data penelitian adalah data teks berbahasa Indonesia yang berisi hasil *review* para pelanggan warung kuliner yang sudah pernah mengunjungi tempat tersebut. Pada penelitian ini studi literatur bersumber dari jurnal-jurnal penelitian, buku, serta internet terkait dengan topik penelitian yang dilakukan untuk mendukung keberhasilan dari penelitian yang dilakukan. Data yang telah didapatkan kemudian diklasifikasikan secara manual menjadi 2 kategori sesuai dengan rating yang diberikan oleh pelanggan, yaitu kategori "Bagus" dan kategori "Rata-rata".

3.2. Preprocessing Data

Tahapan ini adalah tahapan yang dilakukan sebelum *dataset* dimasukan kedalm model yang akan dihasilkan sehingga data yang masuk merupakan data yang sesuai dengan model yang akan dihasilkan. Sebelum dilakukannya proses *preprocessing* terhadap data *text* yang telah didapatkan, terlebih dahulu dilakukan kalsifikasi data *text* berdasarkan kategori. Dari seluruh *dataset* yang digunakan, untuk menentukan kategori klasifikasi data, maka ketentuan yang digunakan adalah seperti Tabel 1.

Tabel 1. Kategori Klasifikasi Data

Jumlah Rating (bintang)	Klasifikasi	
	Peringkat	Kategori
*****	Luar biasa	Bagus
****	Sangat bagus	
***	Rata –rata	
**	Buruk	Rata-rata
*	Sangat buruk	

Pada Tabel 1, penentuan jumlah rating yang sesuai dengan teks diklasifikasikan menjadi dua kategori, yaitu "rata-rata" dan "bagus".

Tahapan selanjutnya adalah proses pengolahan *text* menjadi *input* kedalam model yang akan digunakan. Adapun tahapan dalam proses ini diantaranya sebagai berikut:

1) Transform Cases: yaitu tahapan seluruh text yang akan dimasukan kedalam model dirubah menjadi huruf kecil semua.

***2) Tokenize*:** yaitu sebuah proses pemisahan teks menjadi beberapa bagian atau yang disebut juga token dengan batasan spasi dan tanda baca.

3) Filter Tokens (by Length): yaitu pembatasan jumlah minimal dan maksimal karakter. Nilai parameter *filter tokens* untuk penelitian ini di setting adalah *min chars* = 4, dan *mac chars* = 20.

***4) Stopword*:** yaitu proses menghilangkan teks yang bersesuaian dengan teks pada daftar yang terdapat pada *stopword* yang sudah ditentukan, untuk tahapan ini menggunakan *stopword text* yang isisnya adalah teks berbahasa indonesia.

5) Weighting: yaitu proses pembobotan setiap term, pada tahapan ini menggunakan model *Term Frequency-Inverse Document Frequency* atau TF-

IDF. TD-IDF ini adalah sebuah pemberian bobot dengan menggunakan pola *term frequency* atau jumlah *term* dalam setiap dokumen, dan *inverse document frequency* atau *invers* dari jumlah dokumen yang memuat suatu *term*. (Chen dkk., 2016).

3.3. Penentuan Data Training dan Testing

Pada tahapan ini adalah dilakukannya pembagian *dataset* yang akan digunakan, jumlah dataset yang didapatkan kemudian data dokumen terbagi menjadi data *training* dan data *testing*. *Dataset* dokumen dari jumlah keseluruhan di *split* menjadi 90% ini digunakan sebagai *data training*, dan sisanya yaitu 10% digunakan untuk data testing karena menggunakan *cross validation* (Van der Gaag dkk., 2006). Data dokumen yang digunakan adalah sejumlah 80 dokumen teks yang berisi komentar-komentar dengan ditentukan 39 data masuk dalam kategori "Bagus", dan 40 data lagi masuk dalam kategori "Rata-rata".

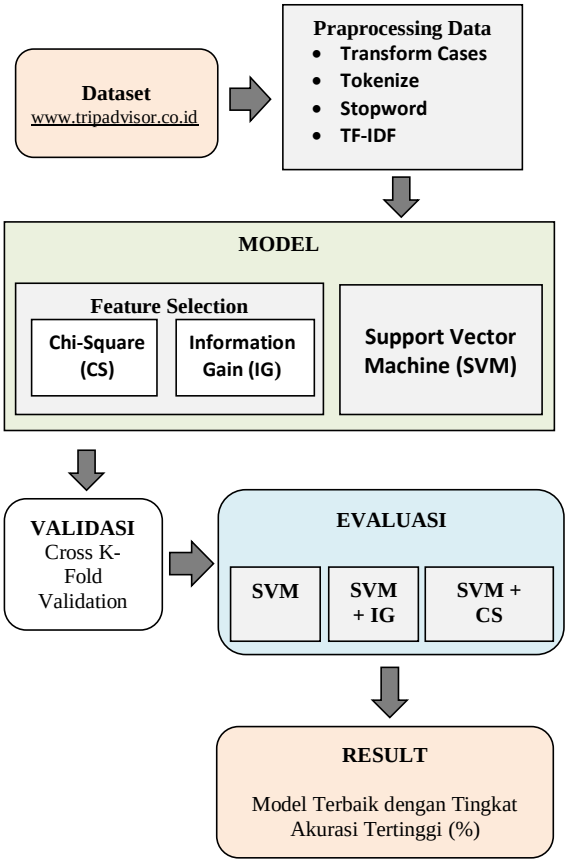
Tabel 2. Contoh dataset teks training dan testing

Nama warung	Training dan testing	kategori
batibul	"Bila anda di tegal dapat mampir ke restoran ini karena lezat dan empuknya sate kambing muda ini meski harganya cukup mahal namun akan terbayar dengan cita rasanya".	Bagus
batibul	"Mantap kambingnya gak berbau.... cocok buat semua keluarga jadinya Ditutup dengan sop nya Dan dengan minum teh tawar angat Segerr	Bagus
sotosedap	"Rasanya enak dan unik saya sampai nambah makannya. Kuahnya dan campuran tauco penambah nikmat rasa sotonya".	Bagus
sateayammargasari	"Lebaran 2017 waktunya makan bersama keluarga nah warung ini kebiasaan dan langganan keluarga saya semuanya terasa nikmat dan sedap"	Rata-rata
sototaucomadi	"Karena penasaran maka saya ingin mencoba. Ternyata enak. Rasanya seperti tom yum ada asam dan manisnya.Tapi baunya tauco.Bisa pilih daging ayam, daging sapi atau babat."	Rata-rata

3.4. Metode Yang Diusulkan

Pada tahapan ini adalah dilakukannya proses ekperimen untuk mencari model terbaik yang diinginkan sesuai dengan *dataset* yang telah diperoleh sebelumnya. Pada penelitian ini diusulkan sebuah metode untuk menghasilkan model terbaik untuk mengukur tingkat kepuasan pelanggan terhadap pelayanan warung dan restoran kuliner di Kota Tegal. Metode pada penelitian ini mengusulkan sebuh metode dengan menggunakan algoritma *Suppor Vector Machine* (SVM) berbasis *feature selection*, diharapkan model yang nantinya dihasilkan adalah model terbaik serta memiliki tingkat akurasi yang lebih baik, digambarkan pada Gambar 2.

Proses pelaksanaan evaluasi model diperlihatkan pada Gambar 2, model yang telah diperoleh akan di evaluasi dengan cara di komparasi dengan model yang lain yaitu SVM klasik dan SVM dengan *feature eslection* yaitu *SVM-IG* dan *SVM-CS*. Setelah dilakukan komparasi diharapkan mendapatkan model terbaik dengan tingkat akurasi yang tertinggi. Pada tahapan analisis data, data yang telah diperoleh dilakukan analisis dengan memasukannya kedalam model yang dihasilkan untuk memperoleh model yang terbaik. Pada tahapan ini, ekperimen dilakukan terus-menerus untuk mendapatkan hasil yang terbaik dengan menggunakan *tools RapidMiner Studio* sebagai pendukung untuk mendapatkan hasil penelitian yang diharapkan.



Gambar 2. Rancangan penelitian yang diusulkan

3.5. Validasi dan Evaluasi Sistem

Pada tahapan ini dilakukan untuk mengetahui model yang diusulkan sesuai dengan yang diharapkan maka dilakukan proses validasi, validasi model pada penelitian ini menggunakan *Cross K-Fold Validation* untuk mengetahui nilai akurasi yang dihasilkan. Setelah didapatkan nilai akurasi yang diharapkan maka dilakukan evaluasi yaitu dengan cara membantingkan tingkat akurasi yang dihasilkan oleh

model lain yaitu SVM tanpa *Feature selection*, dengan model SVM + *feature selection.*

4. HASIL DAN PEMBAHASAN

Setelah dilakukan ekperimen untuk mencari model terbaik dengan tingkat akurasi yang lebih baik terhadap sentiment analisis kepuasan pelanggan terhadap warung dan restoran kuliner di Kota Tegal, maka didapatkan beberapa hasil penelitian pada ekperimen tersebut. Pada penelitian ini menggunakan *tools Rapidminer studio* sebagai software untuk analisis data, dengan Sistem Operasi menggunakan SO Windows7 32Bit, dengan processor Core i5, dan memori 4Gb sebagai pendukung dalam melakukan eksperimen.

4.1. Eksperimen Support Vector Machine (SVM)

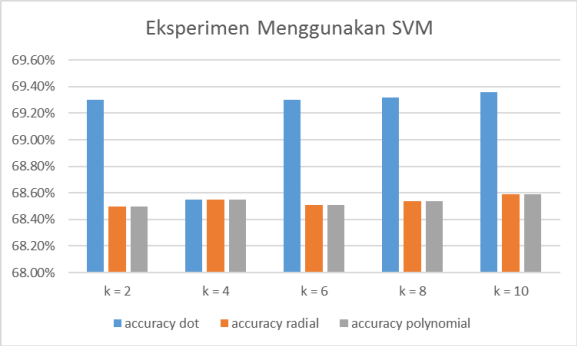
Dengan menggunakan dataset yang telah diperoleh, ekperimen dilakukan untuk mencari model terbaik dengn menggunakan *Support Vector Machine* (SVM). Menentukan nilai parameter SVM pada penelitian ini dilakukan secara manual berbeda dengan adanya optimasi nilai parameter (Friedrichs and Igel, 2005). Sebelum dilakukannya ekserimen, parameter SVM di-*setting* sebagai berikut:

- Tipe *kernel*= dot, radial, polynomial;
- *Kernel cache = 200*;
- C = 0;
- *Convengence epsilon* = 0.001;
- *Nilai max iterations* = 100000;

Hasil eksperimen didapatkan dan diperlihatkan pada Tabel 3.

Tabel 3. Hasil eksperimen SVM

k-Fold	accuracy		
	dot	radial	polynomial
k-Fold = 2	69.30%	68.50%	68.50%
k-Fold = 4	68.55%	68.55%	68.55%
k-Fold = 6	69.30%	68.51%	68.51%
k-Fold = 8	69.32%	68.54%	68.54%
k-Fold = 10	**69.36%**	**68.59%**	**68.59%**



Gambar 3. Rancangan penelitian yang diusulkan

Pada Tabel 2 diperlihatkan hasil eksperimen yang dilakukan dengan menerapkan SVM sebagai algoritma yang digunakan, menghasilkan model dengan tingkat akurasi tertinggi adalah SVM dengan menggunakan kernel = *dot* dan k-Fold = 10 yaitu sebesar 69.36%. Selain itu didapatkan pula tingkat akurasi teringgi dengan menggunakan parameter k-fold=10 dan kernel *radial* sebesar 68,59%, dan tertinggi dengan k-fold=10 menggunakan kernel *polynomial* sebesar 68,59%. Hasil yang didapatkan seperti diperlihatkan pada Gambar 3.

Hasil tertinggi yang tampak pada Gambar 3 menunjukan pada k-Fold =10 pada penerapan SVM ini merupakan parameter dengan menghasilkan tingkat akurasi tertinggi yaitu SVM dengan menggunakan kernel = *dot* dan k-Fold = 10 dengan tingkat akurasi sebesar **69.36%.**

4.2. Eksperimen SVM & Information Gain (IG)

Setelah didapatkannya model terbaik yang didapatkan dengan menggunakan SVM, selanjutnya adalah melakukan eksperimen dengan menerapkan *feature selection* untuk mengoptimalisasi tingkat akurasi klasifikasi yang telah diperoleh oleh SVM. Pada tahapan eksperimen ini dilakukan dengan menggunakan *Information Gain* (IG). Pemilihan atribut *k* dipilih dengan bobot tertinggi (top *k*), nilai parameter *k* yang ditentukan = 10, dan hasil yang didapatkan pada eksperimen ini diperlihatkan pada Tabel 4 dan Tabel 5, serta Tabel 6 dan Tabel 7, selain itu diperlihatkan pula pada Tabel 8.

Tabel 4. Hasil eksperimen SVM-IG dengan K-Fold=10

kernel	accuracy		
	linear sampling	Shuffled sampling	Statified sampling
dot	62.82%	70.83%	**72.44%**
radial	54.87%	66.60%	69.42%
polynomial	66.79%	68.27%	70.19%

Tabel 5. Hasil eksperimen SVM-IG dengan K-Fold=8

kernel	accuracy		
	linear sampling	Shuffled sampling	Statified sampling
dot	61.72%	70.05%	**70.21%**
radial	70.10%	64.53%	70.10%
polynomial	67.19%	69.17%	70.10%

Tabel 6. Hasil eksperimen SVM-IG dengan K-Fold=6

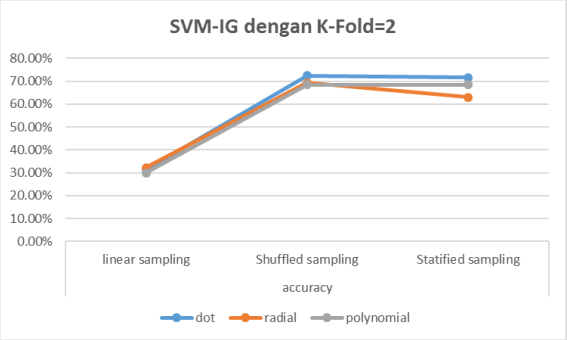
kernel	accuracy		
	linear sampling	Shuffled sampling	Statified sampling
dot	58.04%	71.61%	**71.65%**
radial	52.49%	66.88%	66.16%
polynomial	65.12%	68.51%	70.02%

Tabel 7. Hasil eksperimen SVM-IG dengan K-Fold=4

kernel	accuracy		
	linear sampling	Shuffled sampling	Statified sampling
dot	53.00%	70.16%	**71.65%**
radial	58.39%	64.54%	70.87%
polynomial	65.45%	69.18%	70.87%

Tabel 8. Hasil eksperimen SVM-IG dengan K-Fold=2

kernel	accuracy		
	linear sampling	Shuffled sampling	Statified sampling
dot	31.54%	**72.45%**	71.63%
radial	32.32%	69.30%	63.01%
polynomial	29.97%	68.51%	68.49%



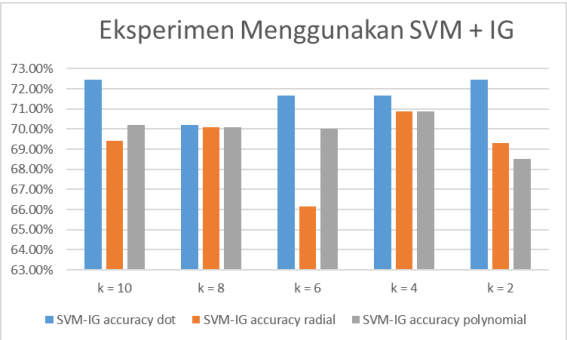
Gambar 4. Rancangan penelitian yang diusulkan

Hasil eksperimen memperlihatkan nilai yang berbeda dari setiap tahapan yang dilakukan, pada Tabel 8 terlihat bahwa tingkat akurasi yang paling tertinggi adalah model SVM-IG dengan tingkat akurasi sebesar 72.45%. PengGambaran grafik atas nilai tertinggi dari model SVM-IG seperti diperlihatkan pada Gambar 4.

Berdasarkan hasil yang didapatkan, untuk nilai akurasi tertinggi seperti tampak pada Tabel 9. Pada Tabel 9 memperlihatkan bahwa tingkat akurasi yang tertinggi untuk model SVM-IG adalah dengan menggunakan *k-Fold* = 2, dan kernel = *dot* yaitu sebesar 72,45%.

Tabel 9. Hasil Eksperimen Model SVM-IG Terbaik

k-Fold	accuracy		
	dot	radial	polynomial
k = 10	72.44%	69.42%	70.19%
k = 8	70.21%	70.10%	70.10%
k = 6	71.65%	66.16%	70.02%
k = 4	71.65%	70.87%	70.87%
k = 2	**72.45%**	69.30%	68.51%



Gambar 5. Rancangan penelitian yang diusulkan

4.3. Eksperimen SVM + Chi Squared Statistic

Untuk optimalisasi tingkat akurasi dengan menggunakan *feature selection* selanjutnya, eksperimen berikutnya dilakukan percobaan dengan menerapkan *Chi Squared* (CS) pada model SVM. Pemilihan atribut *k* dipilih dengan bobot tertinggi (top *k*), nilai parameter *k* yang ditentukan = 10. Hasil penelitian tersebut diperlihatkan seperti pada Tabel 10 dan Tabel 11, serta Tabel 12 dan Tabel 13, selain itu pada Tabel 14.

Tabel 10. Hasil eksperimen SVM- Chi Squared dengan K-Fold=10

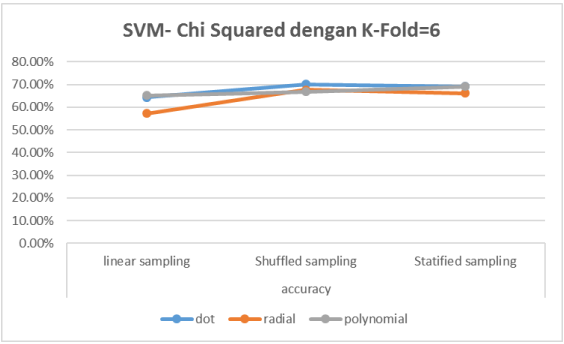
kernel	accuracy		
	linear sampling	Shuffled sampling	Statified sampling
dot	64.36%	69.94%	69.29%
radial	64.36%	65.19%	66.99%
polynomial	64.36%	67.56%	67.76%

Tabel 11. Hasil eksperimen SVM- Chi Squared dengan K-Fold=8

kernel	accuracy		
	linear sampling	Shuffled sampling	Statified sampling
dot	66.35%	66.82%	69.32%
radial	61.46%	65.31%	63.85%
polynomial	64.01%	69.22%	67.71%

Tabel 12. Hasil eksperimen SVM- Chi Squared dengan K-Fold=6

kernel	accuracy		
	linear sampling	Shuffled sampling	Statified sampling
dot	64.32%	**70.09%**	69.23%
radial	57.29%	67.68%	66.13%
polynomial	65.12%	66.92%	69.23%



Gambar 6. Rancangan penelitian yang diusulkan

Pada Gambar 6 memperlihatkan grafik pengGambaran tingkat akurasi tertinggi yang dihasilkan dari model SVM-CS. Hasil eksperimen mendapatkan nilai tertinggi akurasi adalah sebesar 70,09% dengan *type kernel dot* dan menggunakan *Shuffled sampling*. Berdasarkan hasil yang telah didapatkan, maka hasil nilai akurasi tertinggi model SVM-CS adalah seperti ditampilkan pada Tabel 15.

Tabel 13. Hasil eksperimen SVM- Chi Squared dengan K-Fold=4

kernel	accuracy		
	linear sampling	Shuffled sampling	Statified sampling
dot	64.64%	69.23%	69.33%
radial	52.87%	64.52%	66.15%
polynomial	63.84%	66.86%	68.52%

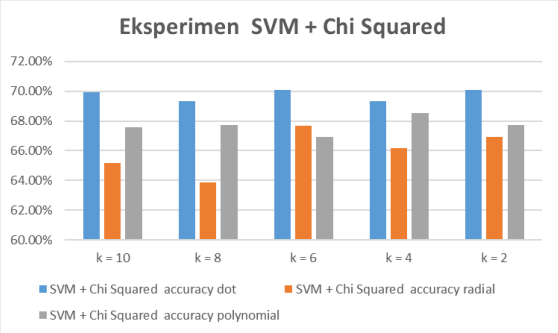
Tabel 14. Hasil eksperimen SVM- Chi Squared dengan K-Fold=2

kernel	accuracy		
	linear sampling	Shuffled sampling	Statified sampling
dot	52.63%	68.50%	70.08%
radial	30.75%	68.51%	66.90%
polynomial	18.25%	66.94%	67.72%

Pada Tabel 15 dan Gambar 7 menunjukan bahwa tingkat akurasi model SVM dan *Chi Squared* (SVM-CS) tertinggi adalah sebesar 70,09% dengan k-Fold = 6 dan *type kernel dot*. Dari hasil yang didapatkan terlihat bahwa tingkat akurasi yang didapatkan dengan menggunakan *type kernel* yang berbeda memiliki tingkat akurasi yang berbeda pula.

Tabel 15. Hasil eksperimen SVM- Chi Squared Tertinggi

k-Fold	accuracy		
	dot	radial	polynomial
k = 10	69.94%	65.19%	67.56%
k = 8	69.32%	63.85%	67.71%
k = 6	**70.09%**	67.68%	66.92%
k = 4	69.33%	66.15%	68.52%
k = 2	70.08%	66.90%	67.72%



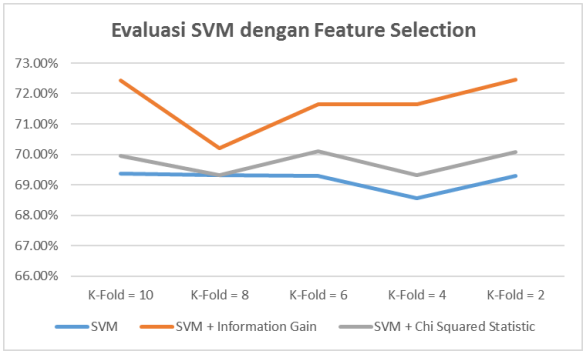
Gambar 7. Rancangan penelitian yang diusulkan

3.2. Evaluasi

Berdasarkan hasil eksperimen yang telah diperoleh, maka dihasilkan model dengan tingkat akurasi tertinggi pada setiap tahapan.

Tabel 16. Evaluasi SVM dengan Feature Selection

K-Fold	SVM	SVM + Information Gain	SVM + Chi Squared Statistic
K-Fold = 10	**69.36%**	72.44%	69.94%
K-Fold = 8	69.32%	70.21%	69.32%
K-Fold = 6	69.30%	71.65%	**70.09%**
K-Fold = 4	68.55%	71.65%	69.33%
K-Fold = 2	69.30%	**72.45%**	70.08%



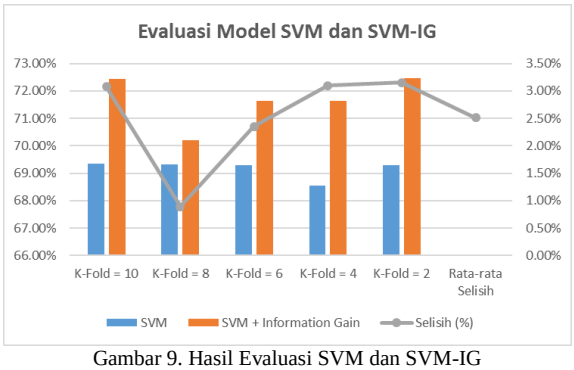
Gambar 8. Hasil evaluasi model yang diusulkan

Berdasarkan hasil penelitian seluruhnya, tampak pada Tabel 16 hasil nilai tingkat akurasi yang dihasilkan, tampak bahwa nilai akurasi tertinggi untuk SVM adalah sebesar 69,36% dengan k-Fold = 10, untuk model SVM + *Information Gain* adalah sebesar 72,45% dengan k=Fold = 2, sedangkan tingkat akurasi dengan model SVM + *Chi Squared* adalah sebesar 70,09%.

Dari hasil yang didapatkan, pada Tabel 17 diperlihatkan evaluasi perbedaan dari tingkat akurasi yang didapatkan dari perbandingan model SVM klasik dan SVM-IG. Untuk perbandingan tingkat akurasi SVM dan SVM-CS ditunjukan pada Tabel 18 dan Gambar 9.

Tabel 17. Evaluasi SVM dan SVM-IG

K-Fold	SVM	SVM + Information Gain	Selisih (%)
K-Fold = 10	69.36%	72.44%	3.08%
K-Fold = 8	69.32%	70.21%	0.89%
K-Fold = 6	69.30%	71.65%	2.35%
K-Fold = 4	68.55%	71.65%	3.10%
K-Fold = 2	69.30%	72.45%	3.15%
Rata-rata Selisih			**2.514%**



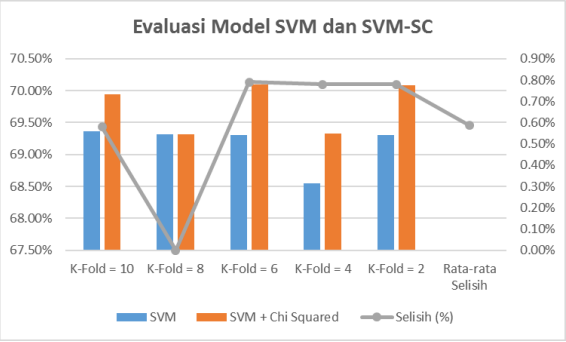
Gambar 9. Hasil Evaluasi SVM dan SVM-IG

Pada Tabel 16 dan Gambar 10 diperlihatkan bahwa tingkat akurasi yang sebelumnya didapatkan oleh SVM kemudian dioptimasi dengan menggunakan *feature selection* SVM-IG dengan rata-rata selisih kenaikan tingkat akurasi sebesar 2.514%. Hal yang sama pada Tabel 18 memperlihatkan adanya keanikan tingkat akurasi setelah dilakukan optimasi

dengan menggunakan model SVM-SC yaitu dengan kenaikan rata-rata sebesar 0.586%.

Tabel 18. Evaluasi SVM dan SVM-CS

K-Fold	SVM	SVM + Chi Squared	Selisih (%)
K-Fold = 10	69.36%	69.94%	0.58%
K-Fold = 8	69.32%	69.32%	0%
K-Fold = 6	69.30%	70.09%	0.79%
K-Fold = 4	68.55%	69.33%	0.78%
K-Fold = 2	69.30%	70.08%	0.78%
Rata-rata Selisih			**0.586%**



Gambar 10. Evaluasi SVM dan SVM-CS

Tabel 19. Hasil Klasifikasi Teks Berdasarkan Model Terbaik

No.	label	prediction (label)	metadata_file
1	Bagus	Bagus	2. batibul-4.txt
2	Bagus	Rata-rata	204.piargasari-4.txt
3	Bagus	Bagus	40. sotosedap-4.txt
4	Bagus	Bagus	71. satekambingtaspirin-4.txt
5	Bagus	Bagus	76. waroengsteak-4.txt
6	Bagus	Bagus	79. waroengsteak-4.txt
7	Bagus	Bagus	90. sototaucomadi-4.txt
8	Rata-rata	Rata-rata	1. batibul-3.txt
9	Rata-rata	Bagus	103.sambellayah-3.txt
10	Rata-rata	Rata-rata	108.sambellayah-3.txt
11	Rata-rata	Rata-rata	198.sambellayah-3.txt
12	Rata-rata	Rata-rata	201.piargasari-3.txt
…	…	…	…
…	…	…	…
123	Bagus	Bagus	54. blengongyanto-4.txt
124	Rata-rata	Bagus	206.pringsewu-1.txt
125	Rata-rata	Bagus	29. batibul-3.txt
126	Rata-rata	Bagus	81. waroengsteak-3.txt
127	Rata-rata	Bagus	93. sototaucomadi-3.txt

Berdasarkan hasil yang telah didapatkan terlihat bahwa optimasi yang dilakukan dengan menggunakan *feature selection* pada SVM mengalami sebuah peningkatan akurasi dan tingkat akurasi yang terbaik adalah dengan menggunakan model SVM-IG.

5. KESIMPULAN

Penerapan *feature selection* dalam pengoptimalisasi tingkat akurasi dalam sentiment analysis klasifikasi rekomendasi pelayanan restoran dan warung kuliner di Kota Tegal telah dapat dilakukan dan telah dapat memberikan sebuah peningkatan akurasi terhadap model SVM yang dihasilkan. *Information Gain* merupakan model yang lebih baik dibandingkan *Chi Squared Statistic* dalam meningkatkan tingkat akurasi SVM, yaitu dengan menghasilkan rata-rata kenaikan tingkat akurasi sebesar 2,514% dengan tingkat akurasi terbaik sebesar 72,45%. Pada penelitian selanjutnya perlu mempertimbangan penelitian lanjutan untuk menemukan model terbaik lagi sehingga tingkat akurasi yang dihasilkan menjadi lebih baik lagi.

UCAPAN TERIMA KASIH

Terimakasih disampaikan kepada DRPM Dirjen Penguatan Riset dan Pengembangan KEMENRISTEK DIKTI DRPM Direktorat Jenderal Penguatan Riset dan Pengembangan Kemenristek DIKTI atas pendanaan penelitian yang diberikan melalui skema pendanaan Penelitian Dosen Pemula (PDP) untuk tahun anggaran 2018.

DAFTAR PUSTAKA

AMINUDIN, A., SN, A. and AHMAD, B., 2018. Automatic Question Generation (AQG) Dari Dokumen Teks Bahasa Indonesia Berdasarkan Non-Factoid Quesion. *Jurnal Teknologi Informasi dan Ilmu Komputer*, [online] 5(2), p.217. Available at: <http://jtiik.ub.ac.id/index.php/jtiik/article/view/664>.

DI CARO, L. and GRELLA, M., 2013. Sentiment analysis via dependency parsing. *Computer Standards and Interfaces*, [online] 35(5), pp.442–453. Available at: <http://dx.doi.org/10.1016/j.csi.2012.10.005>.

CHEN, J., HUANG, H., TIAN, S. and QU, Y., 2009. Feature selection for text classification with Na??ve Bayes. *Expert Systems with Applications*, [online] 36(3 PART 1), pp.5432–5435. Available at: <http://dx.doi.org/10.1016/j.eswa.2008.06.054>.

CHEN, K., ZHANG, Z., LONG, J. and ZHANG, H., 2016. Turning from TF-IDF to TF-IGM for term weighting in text classification. *Expert Systems with Applications*, [online] 66, pp.1339–1351. Available at: <http://dx.doi.org/10.1016/j.eswa.2016.09.009>.

FRIEDRICHS, F. and IGEL, C., 2005. Evolutionary tuning of multiple SVM parameters. *Neurocomputing*, [online] 64, pp.107–117. Available at:

<http://linkinghub.elsevier.com/retrieve/pii/S0925231204005223>.

VAN DER GAAG, M., HOFFMAN, T., REMIJSEN, M., HIJMAN, R., DE HAAN, L., VAN MEIJEL, B., VAN HARTEN, P.N., VALMAGGIA, L., DE HERT, M., CUIJPERS, A. and WIERSMA, D., 2006. The five-factor model of the Positive and Negative Syndrome Scale II: A ten-fold cross-validation of a revised model. *Schizophrenia Research*, 85(1–3), pp.280–287.

JIU-ZHEN LIANG, 2004. SVM multi-classifier and Web document classification. In: *Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.04EX826)*. [online] IEEE, pp.1347–1351. Available at: <http://ieeexplore.ieee.org/document/1381982/>.

KANG, H., YOO, S.J. and HAN, D., 2012. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, [online] 39(5), pp.6000–6010. Available at: <http://dx.doi.org/10.1016/j.eswa.2011.11.107>.

KONCZ, P. and PARALIC, J., 2011. An approach to feature selection for sentiment analysis. In: *2011 15th IEEE International Conference on Intelligent Engineering Systems*. [online] IEEE, pp.357–362. Available at: <http://ieeexplore.ieee.org/document/5954773/>.

LIU, B., 2010. *Sentiment Analysis and Subjectivity*. 2nd ed. [online] Handbook of natural language processing. Available at: <ftp://nozdr.ru/biblio/kolxo3/Cs/CsNl/Indurkhya N., Damerau F.J. (eds.) Handbook of natural language processing (2ed., CRC, 2010)(ISBN 9781420085921)(O)(692s)_CsNl_.pdf#page=653>.

MEDHAT, W., HASSAN, A. and KORASHY, H., 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, [online] 5(4), pp.1093–1113. Available at: <http://dx.doi.org/10.1016/j.asej.2014.04.011>.

MORAES, R., VALIATI, J.F. and GAVIÃO NETO, W.P., 2013. Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, [online] 40(2), pp.621–633. Available at: <http://dx.doi.org/10.1016/j.eswa.2012.07.059>.

RAVI, K. and RAVI, V., 2015. *A survey on opinion mining and sentiment analysis: Tasks, approaches and applications*. [online] *Knowledge-Based Systems*, Elsevier B.V. Available at: <http://dx.doi.org/10.1016/j.knosys.2015.06.015>.

REYES, A. and ROSSO, P., 2012. Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision Support Systems*, [online] 53(4), pp.754–760. Available at: <http://dx.doi.org/10.1016/j.dss.2012.05.027>.

ROBALDO, L. and DI CARO, L., 2013. OpinionMining-ML. *Computer Standards and Interfaces*, 35(5), pp.454–469.

SOMANTRI, O. and KHAMBALI, M., 2017. Feature Selection Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes dan Algoritme Genetika. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI)*, [online] 6(3). Available at: <http://ejnteti.jteti.ugm.ac.id/index.php/JNTETI/article/view/332/257> [Accessed 13 Oct. 2017].

TAN, S. and ZHANG, J., 2008. An empirical study of sentiment analysis for chinese documents. *Expert Systems with Applications*, [online] 34(4), pp.2622–2629. Available at: <http://linkinghub.elsevier.com/retrieve/pii/S0957417407001534>.

TRIPADVISOR LLC, 2017. *10 Restoran Terbaik di Tegal - TripAdvisor*. [online] Available at: <https://www.tripadvisor.co.id/Restaurants-g790289-Tegal_Central_Java_Java.html> [Accessed 17 Oct. 2017].

TRIPATHY, A., AGRAWAL, A. and RATH, S.K., 2015. Classification of Sentimental Reviews Using Machine Learning Techniques. *Procedia Computer Science*, [online] 57, pp.821–829. Available at: <http://dx.doi.org/10.1016/j.procs.2015.07.523>.

WANG, S., LI, D., ZHAO, L. and ZHANG, J., 2013. Sample cutting method for imbalanced text sentiment classification based on BRC. *Knowledge-Based Systems*, [online] 37, pp.451–461. Available at: <http://dx.doi.org/10.1016/j.knosys.2012.09.003>.

WIJOYO, S.H., HERLAMBANG, A.D., ROZI, F. and ISANTA, S.A., 2017. Optimasi Suffix Tree Clustering dengan Wordnet dan Named Entity Recognition untuk Pengelompokan Dokumen. *Jurnal Teknologi Informasi dan Ilmu Komputer*, [online] 4(4), p.263. Available at: <http://jtiik.ub.ac.id/index.php/jtiik/article/view/400>.

ZHANG, Z., YE, Q., ZHANG, Z. and LI, Y., 2011. Sentiment classification of Internet restaurant reviews written in Cantonese. *Expert Systems*

with Applications, [online] 38(6), pp.7674–7682. Available at: <http://dx.doi.org/10.1016/j.eswa.2010.12.147>.

Halaman ini sengaja dikosongkan