Analisis Sentimen pada Kuisioner Kepuasan Terhadap Layanan dan Fasilitas Kampus Universitas Dengan Menggunakan Klasifikasi *Support Vector Machine* (SVM)

Iman Nur Fakhri[1], Jondri[2], Rian Febrian Umbara[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]imannurfakhri@students.telkomuniversity.ac.id, [2]jondri@telkomuniversity.ac.id,
[3]rianfebrianumbara@telkomuniversity.ac.id

Abstrak

Perguruan tinggi sebagai sarana untuk proses memajukan kehidupan berbangsa dan bernegara perlu melakukan adanya peningkatan mutu dan kualitas layanan yang diberikan kepada mahasiswa. Kepuasan mahasiswa dianggap sebagai salah satu masalah utama perguruan tinggi yang harus dipecahkan agar terciptanya perguruan tinggi yang mampu menduduki peringkat nasional maupun internasional. Layanan yang berpengaruh cukup besar dalam hal ini adalah layanan akademik. Tingkat kepuasan mahasiswa terhadap layanan berorientasi pada tenaga pendidik (dosen) sebagai pemberi jasa dan kualitas layanan dalam sarana dan prasarana kegiatan perkuliahan. Oleh sebab itu dilakukan penelitian yang bertujuan untuk mengetahui kepuasan mahasiswa terhadap layanan dan fasilitas universitas. Penelitian dilakukan dengan pengklasifikasian menggunakan *Support Vector Machine* (SVM) dan pembobotan menggunakan *Term Frequency – Invers Document Frequency* (TF-IDF) serta performansi sistem diukur menggunakan Confusion Matrix. Hasil survei kepuasan mahasiswa di Universitas Telkom dari 10000 isian terdapat 67% sentimeen positif dan 33% sentimen negatif. Nilai akurasi tertinggi yang didapatkan pada sistem ini sebesar 70.39% dengan 10000 isian menggunakan kernel Linier.

Kata kunci : Layanan Universitas, Analisis Sentimen, *Support Vector Machine* (SVM), TF-IDF

Abstract

Higher education as a means for the process of advancing the life of the nation and state needs to make an increase in the quality and quality of services provided to students. Student satisfaction is considered as one of the main problems of higher education that must be solved so that universities can be ranked nationally and internationally. Services that have a significant influence in this regard are academic services. The level of student satisfaction with services is oriented to the teaching staff (lecturers) as service providers and service quality in the facilities and infrastructure of lecture activities. Therefore research is conducted which aims to determine student satisfaction with university services and facilities. The research was conducted by classifying using Support Vector Machine (SVM) and weighting using Term Frequency - Inverse Document Frequency (TF-IDF) and system performance measured using Confusion Matrix. The results of the student satisfaction survey at Telkom University from 10000 entries contained 67% positive sentimeen and 33% negative sentiment. The highest accuracy value obtained in this system is 70.39% with 10000 entries using the Linear kernel.

Keywords: University Services, Sentiment Analysis , Support Vector Machine (SVM), TF-IDF

1. Pendahuluan

Data yang diterima oleh sistem informasi Perguruan Tinggi terkait layanan universitas dari kuisioner kepuasan terhadap layanan dan fasilitas mahasiswa sangatlah banyak. Ada dua tipe data yang dikelola yaitu pilihan ganda dan isian teks bebas. Teks bebas merupakan hal yang akan dibahas untuk menetukan informasi yang diterima dari banyaknya mahasiswa yang memberikan komentar kepuasan karena selama ini data dari isian teks bebas hanya dilhat tanpa diketahui seberapa besar hasil tanggapan yang positif atau negatif.

Analisis sentimen adalah studi komputasi dari opini-opini, sentimen, serta emosi yang diekspresikan dalam teks [1]. Tugas dasar dalam analisis sentimen adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, kalimat, atau pendapat. Polaritas mempunyai arti apakah teks yang ada dalam dokumen, kalimat, atau pendapat memiliki aspek positif atau negatif. Salah satu teknik pembelajaran mesin untuk analisis sentimen adalah *Support Vector Machine* (SVM) [2].

Menurut Cristella (2018), yang melakukan analisis sentiment terhadap layanan akademik menggunakan klasisfikasi naïve bayes (kasus : Telkom University), didapatkan hasil akurasi 72%. Pemilihan metode klasifikasi *Support Vector Machine* karena memiliki kemampuan generalisasi dalam mengklasifikasikan suatu *pattern*, tidak termasuk data yang dipakai dalam fase pembelajaran metode tersebut [3].

Dari penelitian yang sudah dilakukan Cristella (2017), yang berjudul Analisis Sentimen Tentang Kasus Layanan Akademik Di Perguruan Tinggi dapat disimpulkan bahwa sentimen analisis dengan menggunakan metode klasifikasi NBC dapat dikatakan efektif dengan hasil akurasi yang dihasilkan kategori baik maka dapat disimpulkan bahwa dengan metode ini dapat membantu para staff dalam menghasilkan/mengetahui tanggapan mahasiswa pada layanan akademik dan hal ini dapat lebih membantu dalam mengevaluasi sistem secara otomatis guna menghasilkan hasil data yang lebih efektif pada layanan akademik di perguruan tinggi untuk tidak menggunakan perhitungan manual, maka pada laporan tugas akhir ini akan digunakan teknik *Preprocessing* yang sama yaitu *Case Folding, Tokenization, Filtering Stopword*, dan *Stemming* kemudian Ekstraksi menggunakan metode N-gram dan Pembobotan menggunakan metode TF-IDF akan tetapi pada laporan tugas ahir ini untuk tahap klasifikasi akan menggunakan metode *Support Vector Machine* (SVM).

Pada penelitian ini, masalah yang diambil adalah bagaimana mengimplementasikan klasifikasi support vector machine (SVM) untuk analisis sentimen pada kepuasan terhadap layanan dan fasilitas universitas dan bagaimana mengukur performansi dari sistem yang telah dibuat. Adapun tujuan dari penelitian ini adalah menganalisis sentimen pada kuisioner kepuasan terhadap layanan dan fasilitas universitas dan mengetahui tingkat performansi analisis sentimen menggunakan metode *Confusion Matrix*. Batasan masalah pada penelitian ini adalah data yang digunakan menggunakan data kuisioner kepuasan terhadap layanan dan fasilitas universitas, serta hasil pengklasifikasian oleh sistem hanya terdapat dua kelas yaitu positif dan negatif.

2. Studi Terkait

2.1 Text Mining

Text mining merupakan teknologi yang digunakan untuk menganalisis data tak terstruktur data berbentuk teks. Dalam analisis *text mining* terdapat dua fase utama yaitu (1) *Preprocessing* dan integrasi dari *data* tak terstruktur, (2) Analisis statistik data yang telah dilakukan *preprocessing* untuk mengekstraksi konten dari yang terdapat dalam teks menurut Francis dan Flynn [6]. Sedangkan menurut *Shollow Weiss* dalam bukunya menyatakan bahwa *text mining* merupakan transformasi dari data teks menjadi data numerik, dengan kata lain *text mining* mengubah data tak terstruktur menjadi data terstruktur [7].

Text mining merupakan sebuah teknik untuk mencari pengetahuan yang berharga dari data teks yang tidak beraturan, termasuk di dalamnya adalah metode *Natural Language Processing* (NLP), *Data Mining*, dan *Information Visualizaion* yang sangat membantu untuk menemukan pengetahuan baru. Dalam *text mining* sendiri, banyak teknik yang dikembangkan dengan tujuan untuk mempermudah proses analisis diantaranya *Information Extraction from Text Data, Text Summarization, Unsupervised Learning Methods from Text Data, LSI and Dimensionality Reduction for Text Mining, Supervied Learning Methods for Text Data, Transfer Learning with Text Data, Probabilistic Technique for Text Mining, Mining Text Streams, Cross-Lingual Mining of Text Data, Text Mining in Multimedia Networks, Text Mining in Social Media, Sentimenon Mining from Text Data, Text Mining from Biomedical Data* berdasarkan *An Introduction to Text Mining* [10].

2.2 Analisis Sentimen

Analisis sentimen merupakan sebuah bidang study yang menganalisis pendapat, sentimen, evaluasi, sikap dan emosi terhadap suatu entitas seperti produk, jasa, organisasi, individu, masalah, topik dan atribut dari entitas tersebut. Terdapat berbagai sebutan dari analisis sentimen ini, seperti sentiment on mining, sentimen on extraction, sentimen mining, subjectivity analysis, affect analysis, emotion analysis, review mining dan lain sebagainya menurut Liu pada penelitiannya tentang Sentiment Analysis and Opinion Mining [1].

2.3 Preprocessing

Preprocessing merupakan tahap yang dilakukan sebelum melakukan tahap pengklasifikasian. Sebelumnya dataset mentah dibersihkan dahulu. Tahap ini dilakukan untuk mempermudah proses pengklasifikasian [9]. Dataset yang dipilih adalah kepuasan layanan akademik dengan kata kunci masukan tertentu.

Data *preprocessing* merupakan tahap yang dilakukan sebelum pengklasifikasian opini. Berikut tahapan yang dilakukan pada *preprocessing*, yaitu:

a.*Case Folding*

Proses ini merubah huruf besar menjadi huruf kecil dan menghilangkan seluruh tanda baca pada kalimat.

b.*Tokenization*

Sebuah dokumen dapat dipecah menjadi bab-bab, bagian, paragraf, kalimat, kata, dan bahkan suku kata dan fonem. Hal tersebut seringkali ditemukan pada sistem text mining yang melibatkan pemotongan teks menjadi kalimat dan kumpulan kata, yang biasa disebut dengan tokenisasi [10].

c.*Filtering Stopword*

Stopword atau *Stopword Removal* adalah proses penghilangan kata-kata yang tidak berkontribusi banyak pada isi dokumen. Kata-kata yang dianggap tidak digunakan dan tidak berpengaruh terhadap isi dokumen akan dihapus atau dihilangkan [9].

Proses membersihkan data dari sebuah karakter-karakter atau kata yang tidak berguna, seperti tanda baca dan preposisi.

Term yang diperoleh dari tahap tokenisasi dicek dalam suatu daftar *stopword*, apabila sebuahkata masuk di dalam daftar *stopword* maka kata tersebut tidak akan diproses lebih lanjut. Sebaliknya apabila sebuah kata tidak termasuk di dalam daftar *stopword* maka kata tersebut akan masuk ke proses berikutnya. Daftar *stopword* tersimpan dalam suatu tabel, yang berisi kata-kata seperti ; ini, itu, yang, ke, di, dalam, kepada, dan seterusnya.

Contoh Ilustrasi :

"Semoga lebih bisa diperbaiki lagi yang masih kurang"

"Semoga lebih diperbaiki yang kurang"

Karena pada daftar stopword ada kata "lagi" maka pada contoh kalimat diatas kata"lagi" dihilangkan. *Stopword* yang digunakan adalah dari *Library Python Sastrawi* [2].

d.*Stemming*

Proses pembersihan sebuah kalimat atau proses mengubah kata berimbuhan menjadi kata dasar, seperti jadi, menjadi, menjadikan.

Stemming ini menggunakan *Library Python Sastrawi*. Algoritma Nazief dan Adriani Tahapan *stemming* meliputi :

1)Langkah pertama adalah memeriksa apakah kata tersebut merupakan akar kata (*root*) terdapat dalam daftar akar kata (*root*). Kita kata tersebut merupakan akar kata, maka proses dihentikan pada tahap pertama ini

2)Menghilangkan *Inflection Suffixes* ("-lah", "-kah", "-ku", "-mu", atau "-nya"). Jika kata berupa *particles* ("-lah", "-kah", "-tah" atau "-pun") maka langkah ini diulangi lagi untuk menghapus *Possesive Pronouns* ("-ku", "-mu", atau "-nya").

3)Menghilangkan *derivational suffix* (imbuhan turunan). Hilangkan imbuhan -i, -kan, -an.

4)Menghilangkan *derivational prefix* (awalan turunan). Hilangkan awalan be-, di-, ke-, me-, pe-, se- dan te-.

5)Bila dari langkah 4 di atas belum ketemu juga. Maka lakukan analisis apakah kata tersebut masuk dalam tabel diambiguitas kolom terakhir atau tidak.

6)Bila semua proses di atas gagal, maka algoritma mengembalikan kata aslinya.

2.4 Pembobotan Fitur (TF-IDF)

Klasifikasi dengan menggunakan SVM dinyatakan dalam bentuk model *vector-space*. Dalam penelitian ini kata-kata akan diproses dalam metode SVM sebelumnya harus diubah terlebih dahulu keadaan bentuk *vector-space*, salah satu caranya dengan melakukan proses pembobotan kata [9].

TF-IDF merupakan metode pembobotan term yang banyak digunakan sebagai metode pembanding terhadap metode pembobotan baru. Metode ini akan menghitung nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) pada setiap token t di dokumen d dengan persamaan berikut :

$$IDF = \log\left(\frac{1+N}{1+dft}\right) + 1 \quad (2\text{-}1)$$

Dimana :

N = jumlah koleksi dokumen

dft = jumlah dokumen dimana terdapat *term* (t)

Jadi pada metode pembobotan TF-IDF, perhitungan bobot *term* dalam sebuah dokumen dilakukan dengan menghilangkan nilai TF dengan nilai IDF. Persamaan (2-2) merupakan perhitungan bobot *term* (W):

$$w_{d,t} = TF_{d,t} \text{ X } IDF_t \quad \text{(2-2)}$$

Dimana:

W = bobot *term* t terhadap dokumen
TF = frekuensi kemunculan *term* t pada dokumen d
IDF = nilai IDF dari *term* t
d = dokumen ke-d
t = kata *term* ke-t

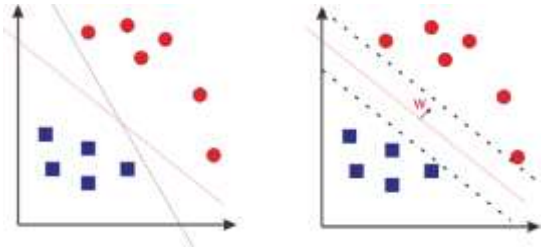
Setelah didapatkan hasil TF-IDF, akan dilakukan normalisasi menggunakan *Euclidean Norm* dengan persamaan (2-3)[11].

$$v_{norm} = \frac{v}{||v||_2} = \frac{v}{\sqrt{{v_1}^2+{v_2}^2+\ldots{v_n}^2}} \quad \text{(2-3)}$$

Dimana, *v* merupakan hasil TF-IDF per dokumen. Bobot *term* akan bernilai semakin besar jika kata sering muncul dalam suatu dokumen dan semakin kecil jika kata tersebut jarang muncul dalam banyak dokumen. Hasil dari pembobotan *term* ini selanjutnya akan digunakan pada tahap aplikasi.

2.5 *Support Vector Machine* (SVM)

Support Vector Machine (SVM) merupakan salah satu teknik klasifikasi dalam data mining. SVM merupakan metode dalam machine learning yang bekerja dengan prinsip *Structural Risk Minimization* (SRM) dengan tujuan menemukan *hyperplane* terbaik yang memisahkan dua buah *class* pada *input space*. Dalam Gambar 2-1.(a) ditunjukkan berbagai alternatif garis pemisah (*discrimination boundaries*). *Hyperplane* pemisah terbaik antara kedua kelas diperoleh dengan cara mengukur *margin* dari *hyperplane* dan mencari *margin* terbesar. *Margin* adalah jarak antara *hyperplane* dengan data terdekat dari setiap kelas. Data terdekat dengan *hyperplane* selanjutnya disebut dengan *support vector*. Dalam penelitian tentang [12] disajikan ilustrasi mengenai *hyperplane* yang akan disajikan dalam Gambar 2-1 di bawah ini dimana alternatif bidang pemisah antara dua kelas (a) dan bidang pemisah terbaik dengan margin terbesar dimana garis solid pada gambar (b) menunjukkan *hyperplane* terbaik.



Gambar 1. Hyperplane terbaik berdasarkan margin optimal

Input dalam SVM merupakan data vektor yang terdiri dari angka *real* [13]. Sedangkan setiap label dinotasikan dengan $y_i \in \{-1, +1\}$ dengan i=1,2,3,…l, dimana l adalah banyaknya data. Dalam [12] diasumsikan kedua class -1 dan +1 dapat terpisah secara sempurna oleh *hyperplane* berdimensi *d*, yang didefinisikan pada persamaan (2-4).

$$\overrightarrow{w}.\vec{x} + b = 0 \quad \text{(2-4)}$$

Dimana:
$\vec{w}$: parameter *hyperplane* yang dicari (garis tegak lurus antara garis *hyperplane* dan titik *support vector*)
$\vec{x}$: data *input* SVM (nilai polaritas dan bobot N-*gram term*) b : parameter *hyperplane*
f : $w^T x + b$

Suatu *pattern* yang termasuk dalam *class* -1 (sampel negatif) dapat dirumuskan melalui pertidaksamaan (2-5). Sedangkan untuk *pattern* dalam *class* +1 (sampel positif) disajikan dalam pertidaksamaan (2-6).

$$\overrightarrow{w}.\vec{x} + b \leq -1 \quad \text{(2-5)}$$

$$\overrightarrow{w}.\vec{x} + b \geq +1 \quad \text{(2-6)}$$

Margin terbesar dapat ditemukan dengan memaksimalkan jarak antara *hyperplane* dengan titik terdekatnya yaitu sebesar $\frac{2}{||\vec{w}||}$. Jika besar $||\vec{w}||$ optimum, maka besar *margin* semakin optimal. Sehingga, optimum $||\vec{w}||$ dirumuskan sebagai *Quadratic Programming* (QP) *problem*, yaitu mencari titik minimal persamaan (2-5), dengan memperhatikan *constraint* persamaan (2-7).

$$\min \tau(w) = \frac{1}{2}||\vec{w}||^2$$

(2-7)

yang memenuhi

$$y_i(\vec{w}.\vec{x} + b) - 1 \geq 0, \forall_i$$

Permasalahan ini dapat dipecahkan dengan berbagai teknik komputasi diantaranya *Lagrange Multiplier* yang ditunjukkan pada persamaan (2-8).

$$L(\vec{w}, b, \alpha) = \frac{1}{2}||\vec{w}||^2 - \sum_{i=1}^{l} \alpha_i [y_i(\vec{w}.\vec{x}_i + b) - 1]$$

(2-8)

Dengan i=1,2,..,l

Dimana:

$\overrightarrow{w}$: parameter *hyperplane* yang dicari (garis tegak lurus antara garis *hyperplane* dan titik *support vector*)

$\vec{x}$: data *input* SVM (nilai polaritas dan bobot N-*gram term*)

b : parameter *hyperplane*

f : $w^T x + b$

y_i : label kelas data *training*

α : variabel *non-negative Lagrange Multiplier*

L : fungsi Lagrangian

α_i merupakan *Lagrange Multiplier* yang bernilai nol atau positif. Solusi dari problem optimisasi dengan pembatas pada persamaan ditentukan dengan *mencari saddle point* dari persamaan (2-8). Maka, persamaan (2-8) diminimalkan terhadap $\overrightarrow{w}$ dan *b*, serta dimaksimalkan terhadap α_i. Dengan memperhatikan sifat bahwa pada titik optimal *gradient L=0*, persamaan (2-9) dapat dimodifikasi sebagai maksimalisasi yang hanya mengandung α_i.

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \vec{x}_i.\vec{x}_j$$

(2-9)

yang memenuhi,

$\alpha_i > 0$, (i = 1,2, … , l) $\sum_{i=l}^{l} \alpha i \, yi = 0$

Dari persamaan (2-9) didapatlah α_i yang kebanyakan bernilai positif. Data yang berkorelasi dengan α_i yang positif inilah yang disebut sebagai *support vector* menurut *Pang, Lee dan Vaithyanathan* dalam penelitiannya pada tahun 2002 [13]. Setelah menemukan *support vector*, maka *hyperplane* pun dapat ditentukan.

Setelah mendapatkan *support vector*, maka nilai $\vec{w}$ dan b dapat dihitung. Nilai $\vec{w}$ dan b akan digunakan sebagai parameter untuk pengujian. Adapun rumus untuk mendapatkan nilai $\vec{w}$ yang merupakan penurunan dari fungsi *Lagrangian* terhadap $\overrightarrow{w}$ yaitu menjadi persamaan (2-10).

$$\vec{w} = \sum_{i=l}^{l} \alpha_i y_i \vec{x}_i$$

(2-10)

Setelah diketahui *support vector* dan nilai $\overrightarrow{w}$, dengan kondisi Karush KuhnTucker (KKT) pada persamaan (2-11):

$$\alpha_i[(y_i(\vec{w}.\vec{x}_i + b) - 1)] = 0$$

(2-11)

Nilai *bias* dapat dicari dengan penyerdehanaan persamaan (2-11) yaitu menjadi persamaan (2-12).

$$b = 1 - \vec{w}.\vec{x}_i$$

(2-12)

Setelah menemukan parameter $\overrightarrow{w}$ dan b, maka fungsi atau model yang digunakan untuk mengklasifikasikan data *testing* terdapat pada persamaan (2-13).

$$f(x) = sign[(\vec{w}.\vec{x}_i) + b]$$

(2-13)

Dimana:

$\overrightarrow{w}$: parameter *hyperplane* yang dicari (garis tegak lurus antara garis *hyperplane* dan titik *support vector*)

$\vec{x}$: data *input* SVM (nilai polaritas dan bobot N-*gram term*)

b : parameter *hyperplane* f : fungsi *hyperplane*

sgn : fungsi *sign* yang membulatkan hasil dari rumus *hyperplane* (+1 atau -1)

Data *input* dapat dipisahkan secara *linear* dengan baik maka disebut *hard margin classification*. Sedangkan pada beberapa data *input* tidak dapat dipisahkan secara *linear* sehingga diberi variabel estimasi kesalahan (ξ_i) disebut *soft margin*. *Slack variable* (ξ_i) digunakan untuk meminimalkan kesalahan klasifikasi

(*misclassification error*) pada data *input*. Dalam *soft margin*, persamaan (2-7) akan dimodifikasi dengan memasukkan *slack variable* ξ_i menjadi persamaan (2-14).

$$min\ \tau(\vec{w},\xi) = \frac{1}{2}\ ||\vec{w}||^2 + c\sum_{i=l}^{l}\xi_i \quad (2\text{-}14)$$

yang memenuhi

$$y_i(\vec{w}.\vec{x}+b) \geq 1-\xi_i, \forall_i$$

Penggunaan variabel untuk mengatasi kasus ketidaklayakan dari pembatas $y_i(\vec{w}.\vec{x_i}+b) \geq 1$ dengan cara memberi penalti untuk data yang tidak memenuhi pembatas tersebut. Untuk meminimalkan nilai ξ_i ini, maka diberikan penalti dengan menerapkan parameter C. Parameter C berfungsi untuk mengontrol *trade off* antara *margin* dan *error* klasifikasi ξ.

2.6 Confusion Matrix

Setelah sistem berhasil dibuat, diperlukan suatu pengukuran atau perhitungan untuk menilai apakah sistem tersebut sudah sesuai dengan yang diharapkan atau belum. Pada penelitian ini akan diadakan suatu penilaian dari keberhasilan sistem berdasarkan keakuratan sistem memprediksi komentar positif atau negatif [9]. Pada Tabel 2.2 menunjukan *confusion matrix* yang biasa digunakan untuk perhitungan dalam

	Predicted Negative	*Predicted Positive*
Actual Negative	*Number of True Negative instances (TN)*	*Number of False Positives instances (FP)*
Actual Positive	*Number of False Negatives instances (FN)*	*Number of True Positives instances (TP)*

pengujian dalam penulisan bahasa Inggris

Tabel 2.2 Confussion Matrix [12]

Dimana:

- TP (*True Positive*) adalah jumlah dokumen komentar positif diprediksi positif oleh sistem.
- FN (*False Negative*) adalah jumlah dokumen komentar positif diprediksi negatif oleh sistem.
- FP (*False Positive*) adalah jumlah dokumen komentar negatif diprediksi positif oleh sistem.
- TN (*True Negative*) adalah jumlah dokumen komentar negatif di prediksi negatif oleh sistem.

Berdasarkan *confusion matrix* tersebut terdapat beberapa parameter yang biasannya digunakan dalam mengukur performansi suatu metode. Parameter yang akan digunakan adalah Precision, Recall, Accuracy.

a. Precision adalah proporsi kasus dengan hal positif yang benar.

$$Precision = \frac{TP}{FP+TP} \quad (2\text{-}15)$$

b. Recall adalah proporsi kasus positif yang diidentifikasi dengan benar.

$$Recall = \frac{TP}{FN+TP} \quad (2\text{-}16)$$

c. Accuracy adalah perbandingan kasus yang diidentifikasi benar dengan jumlah semua kasus.

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \quad (2\text{-}17)$$

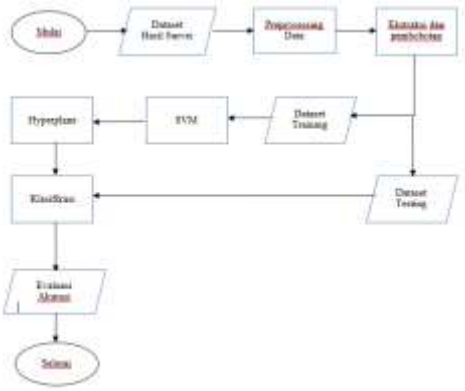
3. Sistem yang Dibangun

Pada bab ini akan dijabarkan secara lengkap mengenai sistem yang dibangun.

3.1 Gambaran Umum

Dalam penelitian ini, dibangun sistem yang dapat menganalisis sentimen terhadap kuisioner kepuasan terhadap layanan dan fasilitas universitas menggunakan klasifikasi SVM. Pada sistem ini, data didapatkan dari data kuisioner kepuasan mahasiswa universitas Telkom. Data diberikan label terlebih dahulu dengan kelas positif dan negatif. Data yang sudah diberikan label akan melewati tahapan preprocessing dan pembobotan bentuk vektor. Data dibagi menjadi dua, yaitu data *training* dan data *testing*. Data *training* yang sudah melewati tahapan *preprocessing* dan pembobotan berbentuk vector akan di *train* menggunakan metode SVM guna mendapatkan hyperplane terbaik yang dapat mengklasifikasikan data menjadi dua kelas yaitu kelas positif dan negatif dari data *training* tersebut. Setelah model yang didapat dari *training* pada data, maka model digunakan sebagai model untuk melakukan *testing*. Berikut adalah gambar dari sistem yang dibangun.

Gambar 3.1. *Flowchart* sistem peneliatian



Berdasarkan gambar 3.1., berikut ini merupakan penjelasan dari tiap proses yang akan dilakukan pada penelitian tugas akhir ini.

1) Dataset

Dataset yang akan dipakai dalam penelitian ini adalah dataset kuisioner kepuasan mahasiswa terhadap layanan dan fasilitas universitas tahun 2017-218 dari sistem Satuan Audit Internal Universitas Telkom. Dari data yang diproses ke dalam sistem, ditunjukkan beberapa sampel komentar mahasiswa terhadap layanan dan fasilitas univeritas pada tabel 3.1.

Tabel 3.1 *sampel komentar mahasiswa terhadap layanan dan fasilitas universitas*

No	Komentar
1	Layanan akademik sudah baik
2	pelayanannya sudah cukup baik
3	tidak usah dipersulit untuk layanan akademik
4	sudah sangat baik pertahankan
5	Sesi kuliah 3 jam melelahkan

2) Preprocessing

Setelah tahap pengumpulan *dataset*, maka tahap selanjutnya adalah *preprocessing* data. *Preprocessing* data merupakan pengolahan pada data asli/mentah untuk dapat digunakan pada proses pengolahan selanjutnya. Beberapa tahapan *preprocessing* yang dilakukan pada penelitian ini yaitu merubah huruf besar menjadi huruf kecil dan menghilangkan seluruh tanda baca pada kalimat (*case folding*), pemecahan kalimat menjadi kata-kata, frase, atau simbol (*tokenization*), membersihan sebuah kalimat atau proses mengubah kata berimbuhan menjadi kata dasar (*Stemming*)

Tabel 3.2 Sampel komentar hasil *Preprocessing* mahasiswa terhadap layanan universitas

No	Komentar	Hasil
1	Layanan akademik sudah baik	layan akademik sudah baik
2	pelayanannya sudah cukup baik	Layan sudah cukup baik
3	tidak usah dipersulit untuk layanan akademik	Tidak usah sulituntuk layan akademik
4	sudah sangat baik pertahankan	Sudah sangat baik tahan
5	Sesi kuliah 3 jam melelahkan	sesi kuliah jam lelah

3) Ekstraksi dan Pembobotan

Ekstraksi dan pembobotan pada penelitian ini digunakan untuk mempersiapkan *dataset* IGRACIAS menjadi fitur – fitur yang akan diolah sebagai data *input classifier*. Hasil dari ekstraksi fitur (*extracted data*) nantinya akan disimpan dalam bentuk vektor dan digunakan sebagai data *input classfier* dan pembobotannya menggunakan TF-IDF. Pembobotan dengan TF-IDF secara tidak langsung telah mewakili penggunaan model *unigram* dengan satu *term* yang terbentuk, sedangkan *bigram* dimanfaatkan agar sistem mampu menangani makna kata frasa. TF menyatakan kemunculan kata dan IDF menunjukkan tingkat kepentingan suatu *term* dalam kumpulan dokumen atau pengukuran keunikan suatu *term* dalam suatu dokumen yang dibandingkan dengan dokumen lain. Pada penelitian ini, proses TF-IDF menggunakan bantuan *library Scikit Learn.* Untuk rumus perhitungan TF-IDF terdapat pada persamaan (2-1) sampai (2-3).

4) Klasifikasi

Klasifikasi sentimen dilakukan dengan pendekatan *supervised learning* menggunakan metode *Support Vector Machine*. Sebelum digunakan dalam proses klasifikasi, *dataset* yang berguna sebagai data input dalam SVM harus diubah ke dalam bentuk vektor dengan ekstraksi fitur. Terdapat dua tahap untuk proses klasifikasinya yaitu proses pembentukan model klasifikasinya dengan memanfaatkan data *training* serta pengujian terhadap model yang telah dibuat dengan memanfaatkan data *testing*.

Pada proses *testing* akan dilakukan pengujian terhadap model klasifikasi yang telah dibuat pada proses *training*. Proses yang dilakukan pada data *testing* sama dengan yang dilakukan pada data *training* seperti *preprocessing* dan ekstraksi fitur kecuali proses klasifikasi. Pada penelitian ini, proses klasifikasi menggunakan bantuan *library Scikit Learn*

5) Evaluasi

Evaluasi performansi dilakukan untuk menguji hasil dari klasifikasi dengan mengukur nilai akurasi dari sistem yang telah dibuat. Pengukuran nilai evaluasi dihitung bedasarkan *input* data *testing* dan model SVM. Evaluasi akan dilakukan dengan menggunakan *confussion matrix. Confusion matrix* adalah tabel yang dibuat untuk menunjukkan tingkat akurasi dari suatu algoritma *machine learning* terutama pada metode *supervised learning* [14]. Semakin tinggi akurasinya, maka semakin baik model yang digunakan dan semakin dapat diandalkan model yang didapat dari penelitian ini.

4. Hasil dan Analisis

4.1 Dataset dan *labeling*

Jumlah dataset yang digunakan sebesar 10000 jumlah komentar. komentar tersebut kemudian dilabelkan secara manual.

Tabel 4.1.1. Pelabelan dari sampel komentar mahasiswa terhadap layanan universitas

No	Komentar	Label
1	Layanan akademik sudah baik	positif
2	pelayanannya sudah cukup baik	positif
3	tidak usah dipersulit untuk layanan akademik	negatif
4	sudah sangat baik pertahankan	positif
5	Sesi kuliah 3 jam melelahkan	negatif

Tabel 4.1.2. menunjukkan jumlah label hasil *labeling* setiap komentar

Tabel 4.1.2. Hasil *labelling*

No	Label	Jumlah
1	Positif	6700
2	Negatif	3300

4.2 Skenario Pengujian

Skenario pengujian dibuat untuk menentukan hasil akurasi terbaik dalam sistem yang telah dibangun. Pada skenario ini digunakan dua buah parameter uji, yaitu, banyaknya data serta data yang terhapus, dan nilai dari jenis kernel SVM .

Skenario dibagi tiga berdasarkan banyaknya data yakni 3000 data, 7000 data, dan 10000 data. Komentar yang melewati *preprocessing* data akan dibagi menggunakan metode *Cross Validation* untuk pemilihan data *training* dan data *testing*.

Kemudian akan dilakukan skenario pengujian pada tabel 4.2.2.

Tabel 4.2.2. Skenario Pengujian

Jumlah data	Jenis Kernel		
	Linear	*Polynomial*	RBF
3000	Skenario 1	Skenario 2	Skenario 3
7000	Skenario 4	Skenario 5	Skenario 6
10000	Skenario 7	Skenario 8	Skenario 9

4.3 Hasil Pengujian

Pengujian dilakukan dengan menggunakan metode *Confusion Matrix* untuk mencari nilai akurasi dari setiap skenario yang telah dibuat. Berikut adalah pengujian data yang dinilai berdasarkan banyaknya data dari semua jenis kernel SVM. Hasil pengujian dapat dilihat pada tabel 4.3.1.

Tabel 4.3.1. Nilai Akurasi Tiap Skenario

Kode Skenario	Akurasi
Skenario 1	63.88%
Skenario 2	65.55%
Skenario 3	65.55%
Skenario 4	67.1%
Skenario 5	67.52%
Skenario 6	67.52%
Skenario 7	70.39%
Skenario 8	69.3%
Skenario 9	69.3%

Dari hasil pengujian diatas maka didapatkan hasil akurasi terbaik pada skenario pengujian pada kode kode skenario 7 dengan 10000 data dan kernel linear sebanyak 70.39%. Hasil pengujian untuk setiap kernel sebgai berikut

Tabel 4.3.2. Pengujian data berdasarkan kernel *Linear*

Banyaknya Data	Akurasi
3000	63.88%
7000	67.1%
10000	69.7%

Tabel 4.3.3. Pengujian data berdasarkan kernel *Polynomial*

Banyaknya Data	Akurasi
3000	65.55%
7000	67.52%
10000	69.3%

Tabel 4.3.4. Pengujian data berdasarkan kernel *RBF*

Banyaknya Data	Akurasi
3000	65.55%
7000	67.52%
10000	69.3%

Tabel 4.3.5 dan 4.3.6 menunjukkan hasil prediksi prediksi dan nilai *Confusion Matrix* dari skenario pengujian terbaik pada kode skenario 7 dengan menggunakan 10000 data dan kernel *Linear*.

Tabel 4.3.5. Tabel Prediksi Confusion Matrix skenario terbaik 10000 data kernel *Linear*

Kelas Sebenarnya	Kelas Prediksi	
	Positif	Negatif
Positif	673	22
Negatif	23	282

Tabel 4.3.6. Tabel Nilai Confusion Matrix skenario terbaik 10000 data kernel *Linear*

Sentimen	*Precision*	*Recall*	*f1-Score*	*Support*
Negatif	0.54	0.08	0.15	308
Positif	0.70	0.97	0.82	695

4.4 Analisis

Dari hasil pengujian yang telah dilakukan, kami mulai dapat menganalisis bahwa banyaknya jumlah data dapat mempengaruhi akurasi dari sistem klasifikasi yang dibuat. Hal itu dapat dilihat dari tabel 4.3.2 pengujian dengan menggunakan 10000 data dan kernel *Linear* yang memiliki tingkat akurasi sebesar 69.3% sedangkan dengan menggunakan 3000 data dan kernel linear hanya menghasilkan akurasi 63.88%.

Dari tabel *confusion matrix* pada tabel 4.3.5, dapat dilihat bahwa kalimat yang berhasil diprediksi dengan baik adalah komentar positif. Hal tersebut dikarenakan banyaknya kumpulan kata pada kelas positif yang terdapat pula pada kelas lainnya.

5. Kesimpulan

Dari hasil penelitian yang telah dilakukan, dapat diambil kesimpulan bahwa sentimen yang paling banyak disampaikan oleh pengisi kuisioner kepuasan terhadap layanan dan fasilitas universitas berdasarkan hasil dari 10000 data yang diujikan adalah 67% sentimen yang bersifat positif. Penggunaan metode SVM untuk melakukan klasifikasi menghasilkan nilai akurasi pada skenario terbaik (10000 data dengan kernel linear) sebesar 69.3%. Kesimpulan dari penelitian ini adalah metode klasifikasi SVM dapat melakukan proses klasifikasi pada analisa sentimen dengan data yang diperoleh pada penelitian ini berupa data yang dapat diklasifikasi secara linear sehingga didapatkan nilai akurasi maksimal menggunakan kernel *linear*.

Daftar Pustaka

[1] Liu, B. (2012). Sentiment Analysis and Sentimenon Mining. Synthesis lectures on human language technologies, 1-167.

[2] Cristella, Elfrida & Sibaroni, Yuliant. 2018. Analisis Sentimen Tentang Kasus Layanan Akademik di Perguruan Tinggi (Kasus : Telkom University). Universitas Telkom. Bandung.

[3] Nugroho, A.S., Witarto, A.B. dan Handoko, D. 2003, Application of Support Vector Machine in Bioinformatics, Proceeding of Indonesian Scientific Meeting in Central Japan, Gifu-Japan, December 20, 2003.

[4] Haryanto, D.J., Muflikhah, L., & Fauzi, M.A. 2018. Analisis Sentimen Review Barang Berbahasa Indonesia Dengan Metode Support Vector Machine Dan Query Expansion. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, Vol. 2, No. 9, pp. 2548-964.

[5] Putranti, D.W., & Winarko, Edi. 2014. Analisis Sentimen Twitter untuk Teks Berbahasa Indonesia dengan Maximum Entropy dan Support Vector Machine. IJCCS, Vol. 8, No. 1, pp. 91-100.

[6] Weis, S. &. (2010). Fundamentals of Predictive Text Mining. Springer-Verlag, Vol.41.

[7] Rosenthal, S., & et.all, &. (2016). Semeval-2016 Task : Sentiment Analysis in Twitter. *Semeval 2016*.

[8] Friedman, J., Hastie., T., & Tibshirani, R. (2001). The elements of Statical Learning. Berlin: Springer Series in Statistics.

[9] Apasari, P.J. 2017. Analisis Sentimen Twitter Menggunakan Metode Lexicon-Based dan Support Vector Machine. Universitas Telkom. Bandung.

[10] Khamair, J. &. (2013). Machine Learning Algorithms for Opinion Mining and Sentiment Classifications. International Journal of Scientific and Research Publications, 1-6.

[11] Scikit Learn. (2017, July 12). Retrieved from Feature Extraction: http://scikit-learn.org/stable/modules/feature_extraction.html#feature-extraction

[12] Han, J., Kamber, M., & Pei, J. (2011). Data Mining Concepts and Tehcniques.

[13] Nugroho, A. &. (2003). Support Vector Machine Teori dan Aplikasinya dalam Bioinformatika. Ilmu Komputer.