

**ANALISIS SENTIMEN BERBASIS ASPEK PADA ULASAN APLIKASI  
TOKOPEDIA MENGGUNAKAN SUPPORT VECTOR MACHINE**

**Skripsi**

**Oleh**

**SABRAH AILIYYA**

**11150940000040**



**PROGRAM STUDI MATEMATIKA**

**FAKULTAS SAINS DAN TEKNOLOGI**

**UNIVERSITAS ISLAM NEGERI SYARIF HIDAYATULLAH**

**JAKARTA**

**2020 M / 1441 H**

**ANALISIS SENTIMEN BERBASIS ASPEK PADA ULASAN  
APLIKASI TOKOPEDIA MENGGUNAKAN SUPPORT VECTOR  
MACHINE**



**PROGRAM STUDI MATEMATIKA**  
**FAKULTAS SAINS DAN TEKNOLOGI**  
**UIN SYARIF HIDAYATULLAH JAKARTA**  
**2020 M / 1440 H**

## PERNYATAAN

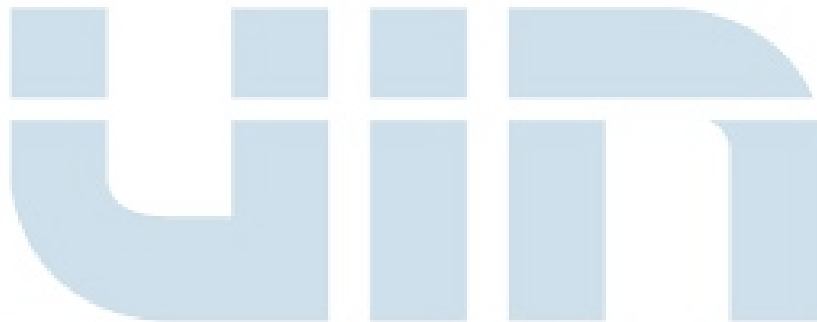
DENGAN INI SAYA MENYATAKAN BAHWA SKRIPSI INI BENAR-BENAR HASIL KARYA SENDIRI YANG BELUM PERNAH DIAJUKAN SEBAGAI SKRIPSI ATAU KARYA ILMIAH PADA PERGURUAN TINGGI ATAU LEMBAGA MANAPUN.

Jakarta, April 2020



Sabrah Ailiyya

11150940000040



## LEMBAR PENGESAHAN

Skripsi berjudul “Analisis Sentimen Berbasis Aspek pada Ulasan Aplikasi Tokopedia Menggunakan Support Vector Machine” yang ditulis oleh **Sabrah Ailiyya**, NIM 11150940000040 telah diuji dan dinyatakan lulus dalam sidang Munaqosyah Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta pada hari Kamis, 9 April 2020. Skripsi ini telah diterima sebagai salah satu syarat untuk memperoleh gelar sarjana strata satu (S1) Program Studi Matematika.

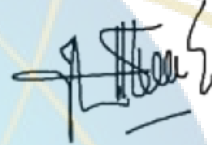
Menyetujui,

Pembimbing I



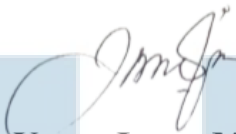
**Dr. Taufik Edy Sutanto, MScTech**  
NIP. 197905302006041002

Pembimbing II



**Dr. Nina Fitriyati, M.Kom**  
NIP. 197604142006042001

Penguji I



**Yanne Irene, M.Si**  
NIP. 197412312005012018

Penguji II

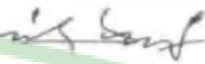


**M. Irvan Septiar Musti, M.Si**  
NUP. 9920113224

Mengetahui,

Dekan Fakultas Sains dan Teknologi

Ketua program Studi matematika



**Prof. Dr. Lily S Eka P, M.Env.Stud**

NIP. 196904042005012005



**Dr. Suma'inna, M.Si**

NIP. 197912082007012015

**LEMBAR PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH  
UNTUK KEPENTINGAN AKADEMIS**

Yang bertanda tangan di bawah ini:

Nama : Sabrah Ailiyya

NIM : 11150940000040

Program Studi : Matematika Fakultas Sains dan Teknologi

Demi pengembangan ilmu pengetahuan, saya menyetujui untuk memberikan **Hak Bebas Royalti Non-Eksklusif** (*Non-Exclusive-Free Right*) kepada Program Studi Matematika Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta atas karya ilmiah saya yang berjudul:

“Analisis Sentimen Berbasis Aspek pada Ulasan Aplikasi Tokopedia Menggunakan Support Vector Machine”

beserta perangkat yang diperlukan (bila ada). Dengan Hak Bebas Royalti Non-Eksklusif ini, Program Studi Matematika Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta berhak menyimpan, mengalihmedia/formatkan, mengelolanya dalam bentuk pangkalan data (*database*), mendistribusikannya, dan menampilkan/mempublikasikannya di internet dan media lain untuk kepentingan akademis tanpa perlu meminta izin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta. Segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta karya ilmiah ini menjadi tanggungjawab saya sebagai penulis.

Demikian pernyataan ini yang saya buat dengan sebenarnya.

Dibuat di Tangerang Selatan

Pada tanggal: 9 April 2020

Yang membuat pernyataan



(Sabrah Ailiyya)

## PERSEMBAHAN DAN MOTTO

Puji dan syukur atas kehadiran Allah SWT dalam hidup, atas izin dan berkatNya penulis mampu menyelesaikan skripsi ini.

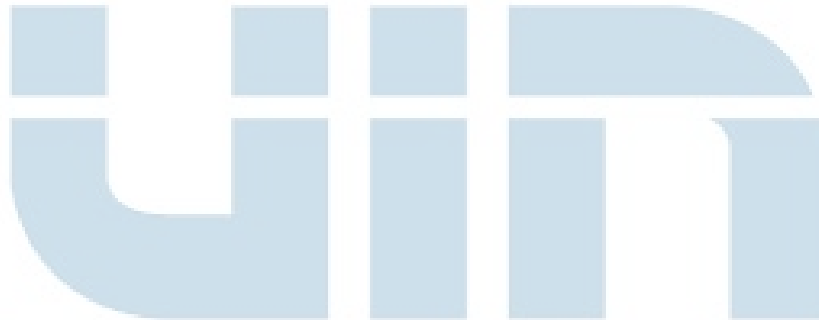
Sholawat beriring salam tercurahkan kepada Nabi Muhammad SAW

Skripsi ini penulis persembahkan untuk kedua orangtua beserta keluarga yang telah memberikan doa serta dukungan yang begitu luar biasa

Skripsi ini juga dipersembahkan untuk yang terkasih, teman-teman.

### MOTTO

“Sabar, satu persatu.”



## ABSTRAK

**Sabrah Ailiyya**, Analisis Sentimen Berbasis Aspek pada Ulasan Aplikasi Tokopedia Menggunakan Support Vector Machine, di bawah bimbingan **Dr. Taufik Edy Sutanto**, MScTech dan **Dr. Nina Fitriyati**, M.Kom.

Pada tahun 2018, sebanyak 171,17 juta penduduk Indonesia merupakan pengguna internet dengan kegiatan yang digemarinya salah satunya adalah melakukan jual dan belanja *online*. Hal tersebut menjadikan banyaknya bermunculan perusahaan dibidang *e-commerce* di Indonesia. Maka dari itu perlu bagi pihak perusahaan untuk mengetahui tentang apa yang menjadi kelebihan serta kekurangan menurut pengguna agar dapat dijadikan sebagai acuan dalam melakukan evaluasi untuk terus meningkatkan kualitas. Karenanya peneliti mengusulkan penelitian *Aspect-Based Sentiment Analysis*, yaitu mengekstrak sentimen dan aspek dari ulasan aplikasi tersebut. Data yang digunakan adalah ulasan pengguna Tokopedia berdasarkan situs *Google Play Store*. Pada penelitian ini dilakukan dua kali klasifikasi, yaitu klasifikasi sentimen dan aspek menggunakan *Support Vector Machine*. Aspek yang digunakan adalah layanan, sistem dan kebermanfaatan. Pemilihan parameter terbaik menggunakan *Grid Search CV* dengan hasil yaitu kernel linear dengan  $c=1$  untuk model klasifikasi sentimen dan aspek. Hasil klasifikasi sentimen dan aspek berturut-turut menunjukkan akurasi sebesar 69,6% dan 74,2%. Dari penelitian ini diperoleh bahwa aspek yang harus diperbaiki oleh Tokopedia adalah layanan.

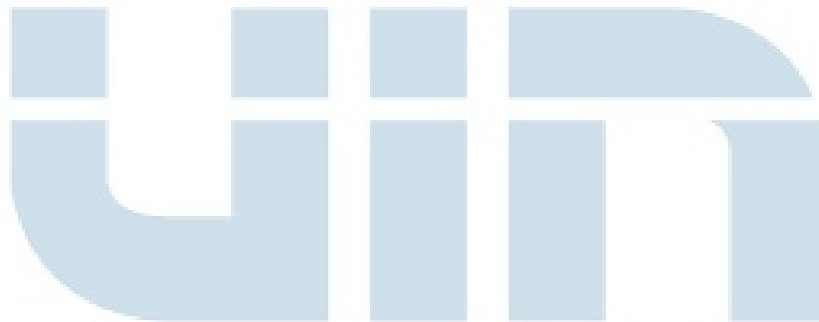
**Kata Kunci:** *Aspect-Based Sentiment Analysis*, *Support Vector Machine*, *Grid Search Cross Validation*.

## ABSTRACT

**Sabrah Ailiyya**, Aspect Based Sentiment Analysis of Tokopedia Application's Review Using Support Vector Machine, under the guidance of **Dr. Taufik Edy Sutanto, MScTech** and **Dr. Nina Fitriyati, M.Kom.**

In 2018, 171.17 million people in Indonesia were internet users with selling and shopping online are things they did the most and it makes the rise of e-commerce in Indonesia. Therefore the company needs to find out about their excellence and weakness according to the user so it can be used as a reference in evaluating to improve the quality. The researcher proposes an Aspect-Based Sentiment Analysis study, which extracts sentiments and aspects from the review of the application. The data used is a Tokopedia user review based on the Google Play Store. In this research, sentiment and aspect classification are using Support Vector Machine. The aspects used are service, system and usefulness. The best parameter selection using Grid Search CV and the result is linear kernel with  $c = 1$  for sentiment and aspect classification models. The result shows the accuracy of sentiment and aspect classification respectively 69.6% and 74.2%. Based on this research, the aspect that Tokopedia has to improve is service.

**Keywords:** Aspect-Based Sentiment Analysis, *Support Vector Machine*, Grid Search Cross Validation.





## KATA PENGANTAR

*Assalamu'alaikum Wr. Wb*

Alhamdulillah, puji dan syukur peneliti panjatkan kepada Allah SWT karena berkat rahmat dan hidayah-Nya penulis dapat menyelesaikan penelitian ini. Shalawat serta salam peneliti curahkan kepada junjungan nabi besar Nabi Muhammad SAW beserta keluarganya, para sahabat dan para pengikutnya.

Peneliti menyelesaikan penelitian ini untuk memperoleh gelar sarjana Matematika. Dalam penyusunan, peneliti tidak luput dari kesulitan dan hambatan. Namun, terdapat pihak – pihak yang memberikan doa, bantuan, motivasi dan selalu menyemangati sehingga penelitian ini dapat terselesaikan. Oleh karena itu peneliti mengucapkan terima kasih kepada:

1. Prof. Dr. Lily Surayya Eka Putri, M.Env.Stud selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Syarif Hidayatullah Jakarta.
2. Ibu Dr. Suma'inna, M.Si, selaku Ketua Program Studi Matematika Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta dan Ibu Irma Fauziah M.Sc, selaku Sekretaris program studi Matematika Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta.
3. Bapak Dr. Taufik Edy Sutanto, M.Sc.Tech selaku pembimbing I dan Ibu Dr. Nina Fitriyati, M.Kom selaku pembimbing II atas ilmu dan arahnya selama penyusunan skripsi ini hingga akhirnya dapat terselesaikan.
4. Ibu Yanne Irene, M.Si selaku penguji I dan Bapak M. Irvan Septiar Musti, M.Si selaku penguji II, terima kasih atas kritik dan sarannya kepada penulis, serta bersedia meluangkan waktunya untuk menguji seminar hasil dan sidang skripsi.
5. Umi dan Ayah, Ka Apit, Teh Noni, Ka Abi, Ka Indah dan Hakim yang tiada hentinya memberikan doa, motivasi dan dukungan hingga peneliti mampu menyelesaikan skripsi ini.

6. Hamid, Ery dan Shinta yang selalu menjadi tempat untuk berkeluh kesah saat menemui kesulitan.
7. Teman – teman yaitu Afifah, Uu, Nunik, Rahil dan Tanjung yang selalu hadir di segala suka dan duka selama kuliah hingga akhir.
8. Teman – teman Matematika 2015 UIN Syarif Hidayatullah Jakarta yang tidak dapat disebutkan satu – persatu.
9. Seluruh pihak yang secara langsung maupun tidak langsung telah membantu, mendukung, serta mendoakan penulis dalam penyelesaian skripsi ini. Meski tidak tertulis namun tidak mengurangi rasa cinta dan terima kasih dari penulis.

Penulis menyadari bahwa masih ada kesalahan dalam penyusunan skripsi ini. Maka dari itu penulis mengharapkan kritik dan saran yang membangun supaya menjadi bahan perbaikan bagi peneliti selanjutnya. Penulis juga berharap penelitian ini bermanfaat bagi siapapun yang membacanya.

*Wassalamu'alaikum Wr. Wb.*

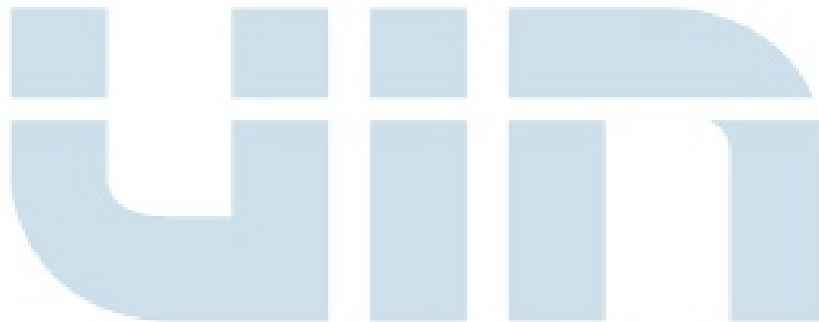
Ciputat, 9 April 2020

Penulis

## DAFTAR ISI

PERNYATAAN.....	iii
LEMBAR PENGESAHAN .....	iv
ABSTRAK .....	vii
ABSTRACT .....	viii
KATA PENGANTAR.....	ix
DAFTAR ISI.....	xi
DAFTAR TABEL .....	xiii
DAFTAR GAMBAR.....	xiv
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang .....	1
1.2 Rumusan Masalah .....	3
1.3 Batasan Masalah.....	4
1.4 Tujuan Penelitian .....	4
1.5 Manfaat Penelitian .....	4
BAB II LANDASAN TEORI .....	5
2.1 <i>Aspect-Based Sentiment Analysis</i> .....	5
2.2 Teknik Pengambilan Data .....	5
2.3 <i>Preprocessing</i> .....	6
2.4 <i>Vector Space Model</i> .....	7
2.5 Norm dan Dot Product.....	8
2.6 Kernel .....	9
2.7 <i>Grid Search Cross Validation</i> .....	11
2.8 Evaluasi Model .....	11
BAB III METODELOGI PENELITIAN.....	13
3.1 Sumber Data .....	13
3.2 Pelabelan Sentimen .....	15
3.3 Pelabelan Aspek.....	16

3.4	Data Training dan Data Testing .....	17
3.5	Support Vector Machine.....	17
3.6	Analisa Numerik SVM.....	22
3.7	<i>Soft Margin</i> .....	25
3.8	Alur Penelitian .....	27
BAB IV HASIL DAN PEMBAHASAN .....		28
4.1	Hasil Preprocessing dan Text Analytics .....	28
4.2	Pembobotan Kata .....	32
4.3	Hyperparameter Tunning .....	33
4.4	Hasil Klasifikasi Sentimen dan Aspek.....	33
4.5	Visualisasi dan Interpretasi Data .....	35
BAB V PENUTUP.....		39
5.1	Kesimpulan .....	39
5.2	Saran .....	40
DAFTAR PUSTAKA.....		41
LAMPIRAN.....		44

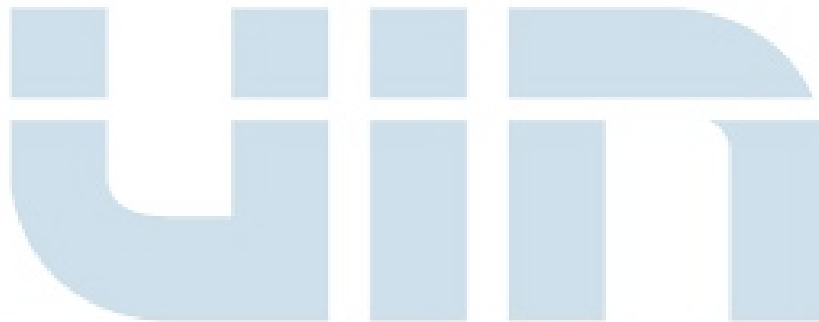


## DAFTAR TABEL

<b>Tabel 2. 1</b> Confusion Matrix .....	12
<b>Tabel 2. 2</b> Rumus Evaluasi Klasifikasi .....	12
<b>Tabel 3. 1</b> Data Awal Ulasan Tokopedia.....	14
<b>Tabel 4. 1</b> Hasil Preprocessing dan Pelabelan.....	28
<b>Tabel 4. 2</b> Hasil Grid Search CV Tiap Kernel Sentimen. ....	33
<b>Tabel 4. 3</b> Hasil Grid Search CV Tiap Kernel Aspek. ....	33
<b>Tabel 4. 4</b> Confusion Matrix Sentimen. ....	34
<b>Tabel 4. 5</b> Confusion Matrix Aspek. ....	34
<b>Tabel 4. 6</b> Positif dan Negatif Tiap Aspek .....	38

## DAFTAR GAMBAR

<b>Gambar 1. 1</b> Penetrasi Pengguna Internet 2018..	1
<b>Gambar 1. 2</b> Peta Layanan E-Commerce Indonesia.....	2
<b>Gambar 3. 1</b> Jumlah Sentimen Tiap Bulan.....	15
<b>Gambar 3. 2</b> Jumlah Aspek Tiap Bulan.....	16
<b>Gambar 3. 3</b> Contoh Hyperplane Dua Dimensi.....	18
<b>Gambar 3. 4</b> Alur Penelitian.....	27
<b>Gambar 4. 1</b> Jumlah Ulasan Tiap Bulan.....	30
<b>Gambar 4. 2</b> Jumlah Sentimen pada Tiap Aspek. ....	30
<b>Gambar 4. 3</b> Wordcloud Label Sentimen (a) Positif dan (b) Negatif.....	31
<b>Gambar 4. 4</b> Wordcloud Aspek (a) Layanan (b) Sistem (c) Kebermanfaatan.....	31
<b>Gambar 4. 5</b> Jumlah Kemunculan Kata.....	32
<b>Gambar 4. 6</b> Wordcloud Aspek Layanan Sentimen (a) Positif dan (b) Negatif.....	35
<b>Gambar 4. 7</b> Wordlink Aspek Layanan Sentimen (a) Positif dan (b) Negatif.....	35
<b>Gambar 4. 8</b> Wordcloud Aspek Sistem Sentimen (a) Positif dan (b) Negatif.....	36
<b>Gambar 4. 9</b> Wordlink Aspek Sistem Sentimen (a) Positif dan Negatif. ....	36
<b>Gambar 4. 10</b> Wordcloud Aspek Kebermanfaatan Sentimen (a) Positif dan (b) Negatif. ....	37
<b>Gambar 4. 11</b> Wordlink Aspek Kebermanfaatan Sentimen (a) Positif dan (b) Negatif. ....	37



## BAB I

### PENDAHULUAN

Pada bab ini penulis akan menjelaskan tentang gambaran umum pelaksanaan penelitian yang mencakup latar belakang, rumusan masalah, batasan masalah, tujuan penelitian dan manfaat penelitian. Hal tersebut akan dijelaskan secara berurutan pada bab ini.

#### 1.1 Latar Belakang

Pesatnya perkembangan teknologi informasi dan digital menjadikan adanya banyak perubahan di hampir seluruh aspek kehidupan. Diantaranya yang terjadi adalah cara manusia berkomunikasi, berdagang dan berbelanja. Oleh sebab itu, banyak perusahaan yang mengubah proses jual-belinya menjadi menggunakan *e-commerce*.

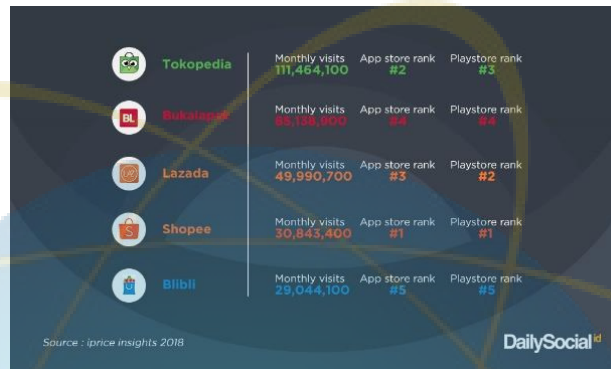
Allah swt. Berfirman: “Katakanlah: *“Hai kaumku, bekerjalah sesuai dengan keadaanmu, sesungguhnya aku akan bekerja (pula), maka kelak kamu akan mengetahui”* (QS Az-Zumar:39). Berdasarkan terjemah ayat tersebut, Allah telah memerintahkan seluruh manusia untuk bekerja sesuai dengan keadaan yang dihadapi. Seperti yang terjadi pada masa sekarang, manusia dalam menjalani kehidupannya adalah dengan menggunakan internet.



**Gambar 1. 1** Penetrasi Pengguna Internet 2018 [1].

Berdasarkan Gambar 1.1, sebanyak 171,17 juta penduduk Indonesia telah menjadi pengguna internet. Kegiatan yang digemari untuk dilakukan ketika

menggunakan internet diantaranya adalah berkomunikasi lewat pesan, menggunakan media sosial, jual dan belanja *online* [1]. Dengan tingginya minat penduduk Indonesia dalam melakukan jual beli secara *online*, menjadikan banyaknya bermunculan perusahaan yang bergerak dibidang perdagangan elektronik atau *e-commerce* di Indonesia.



**Gambar 1. 2** Peta Layanan *E-Commerce* Indonesia [2].

Pada Gambar 1.2 telah disebutkan lima *e-commerce* di Indonesia dengan jumlah kunjungan paling banyak setiap bulannya pada tahun 2018 [2]. Setiap perusahaan tentu telah memiliki strateginya masing-masing untuk tetap bertahan di tengah ketatnya persaingan ini. Tetap perlu bagi pihak perusahaan untuk mengetahui tentang apa yang menjadi kelebihan serta kekurangan menurut pengguna agar dapat dijadikan sebagai acuan dalam melakukan evaluasi untuk terus meningkatkan kualitas. Salah satu upaya yang dilakukan untuk mengetahui hal tersebut adalah dengan dilakukan *Aspect-Based Sentiment Analysis* (ABSA). Untuk penelitian ini, peneliti akan menjadikan Tokopedia sebagai objek. Penelitian semacam ini telah dilakukan sebelumnya oleh Susanti Gojali dan Masayu Leylia Khodra pada tahun 2016 melakukan analisis sentimen berdasarkan aspek dilakukan pada data ulasan restoran [3]. Pada penelitian tersebut, digunakan empat kategori aspek diantaranya adalah *food*, *service*, *price* dan *place*. Data yang digunakan adalah sebanyak 365 ulasan dari 15 restoran pada aplikasi Trip Advisor. Penelitian ini melakukan ekstraksi aspek dan pendapat dari kalimat, serta menentukan orientasi sentimen. Aspek-aspek dikelompokkan bersama dengan menggunakan WordNet dengan pengetahuan sebelumnya tentang kategori aspek. Terdapat



beberapa model yang diujikan diantaranya *Support Vector Machine* (SVM), *Naïve Bayes Classifier* (NBC) dan J48.

Penelitian serupa juga telah dilakukan oleh Puteri Prameswari, Isti Surjandari dan Enrico Laoh di tahun 2017 dengan menggunakan data ulasan pengguna hotel di Bali pada aplikasi Trip Advisor [4]. Penelitian ini menggunakan pendekatan *text mining* dan analisis sentimen berbasis aspek untuk mendapatkan pendapat pengguna hotel dalam bentuk sentimen. Aspek yang digunakan diantaranya *food and beverage operations, accessibility, human resource, activities and entertainments, guests' perspective, transportation service, room amenities* dan *physical environment*. Dengan menggunakan model *Recursive Neural Tensor Network* (RNTN). Hasil dari penelitian tersebut diharapkan dapat digunakan untuk evaluasi dalam meningkatkan kualitas industri perhotelan serta mendukung industri pariwisata di Indonesia.

Hal yang menjadikan penelitian ini berbeda dengan sebelumnya adalah data yang digunakan adalah ulasan dari pengguna aplikasi Tokopedia pada *Google Play Store*. Selain itu, aspek yang digunakan adalah layanan, sistem dan kebermanfaatan. Sedangkan model yang digunakan adalah SVM.

## **1.2 Rumusan Masalah**

Berdasarkan latar belakang penelitian tersebut, maka perumusan masalahnya adalah:

1. Seberapa optimal model ABSA dengan menggunakan metode SVM dalam mengklasifikasikan teks berbahasa Indonesia mengenai ulasan Tokopedia berdasarkan situs *Google Play Store*?
2. Informasi (*insight*) apa saja yang diperoleh dari setiap kelas sentimen pada aspek?

### 1.3 Batasan Masalah

Batasan masalah yang ditentukan untuk menghindari perluasan pembahasan dalam penelitian ini adalah sebagai berikut:

1. Data yang akan diklasifikasi adalah data ulasan pengguna Tokopedia berdasarkan situs *Google Play Store* pada bulan April sampai dengan bulan Juli 2019.
2. Ulasan pada aplikasi yang akan diklasifikasi adalah yang berbahasa Indonesia.
3. Kelas yang digunakan untuk klasifikasi sentimen hanya positif dan negatif.
4. Kelas yang digunakan untuk klasifikasi aspek adalah layanan, system dan kebermanfaatan.

### 1.4 Tujuan Penelitian

Tujuan penelitian ini adalah:

1. Mengetahui seberapa optimal ABSA dengan menggunakan metode SVM dalam mengklasifikasikan teks berbahasa Indonesia mengenai ulasan Tokopedia berdasarkan situs *Google Play Store*.
2. Mengetahui informasi apa yang diperoleh dari setiap kelas sentimen pada aspek.

### 1.5 Manfaat Penelitian

Melalui penelitian ini diharapkan dapat memudahkan Tokopedia dalam mengetahui persepsi pengguna dalam bentuk opini negatif dan opini positif, sehingga dapat dijadikan sebagai acuan dalam upaya menjaga kualitas dan memperbaiki kekurangan serta evaluasi ke arah yang lebih baik.

## **BAB II**

### **LANDASAN TEORI**

Bab ini menjelaskan definisi dan teori-teori yang digunakan sebagai landasan pelaksanaan penelitian yaitu penjelasan tentang ABSA, *preprocessing*, VSM, *cross validation* dan evaluasi model. Teori-teori tersebut dijelaskan secara berurutan pada bab ini.

#### **2.1 Aspect-Based Sentiment Analysis**

Analisis sentimen merupakan suatu analisis berdasarkan teks dengan tujuan untuk mengklasifikasikan teks tersebut berupa opini berdasarkan sentiment [5]. Kemudian para peneliti menemukan *aspect-based sentiment analysis* yang mana dapat melakukan analisis sentimen secara lebih dalam dari suatu teks ulasan. Seperti saat sedang melihat ulasan dari sebuah lagu, opini yang ada tentu bukan hanya sentimen secara keseluruhan melainkan terdapat pula aspek yang spesifik seperti vokal, lirik, kualitas rekaman, dan lainnya [6].

ABSA secara otomatis dapat mengekstraksi aspek-aspek dalam ulasan, kemudian menentukan bagaimana sentimen dari aspek-aspek tersebut [7]. Tahapan pada penelitian ini adalah dengan melakukan dua kali klasifikasi, yaitu klasifikasi aspek dan klasifikasi sentimen.

#### **2.2 Teknik Pengambilan Data**

Dengan adanya perkembangan teknologi, menjadikan para pengguna *e-commerce* mempunyai lebih banyak ruang untuk menyampaikan pengalamannya saat bertransaksi. Salah satu yang dijadikan sebagai wadah untuk menyampaikannya adalah kolom ulasan atau *review* pada Google Play Store yang kemudian dapat dijadikan sebagai bahan untuk penelitian. Melalui ulasan tersebut, akan diperoleh informasi persepsi pengguna dalam bentuk opini negatif dan opini positif yang dapat dijadikan sebagai bahan evaluasi untuk menuju ke arah yang

lebih baik. Untuk memperoleh data ulasan tersebut diperlukan teknik *scraping* yang merupakan teknik untuk mengambil data pada situs *web* secara langsung [8].

### 2.3 *Preprocessing*

Data ulasan dari para pengguna yang diperoleh dari *Google Play Store*, umumnya menggunakan kata yang tidak berstruktur seperti simbol, angka, emotikon dan singkatan sehingga perlu dilakukan *preprocessing*. Berikut adalah *preprocessing* yang dilakukan pada data untuk penelitian ini:

#### 1. *Case Folding*

Pada data ulasan yang digunakan, terdapat banyak penggunaan huruf kapital yang tidak konsisten sehingga *case folding* ini dibutuhkan. Tujuannya adalah untuk menyamakan bentuk pada kata, contohnya adalah seperti yang dilakukan pada data ulasan yang digunakan pada penelitian ini yaitu mengubah semua huruf menjadi huruf kecil atau disebut *lowercase* [9].

#### 2. Menghapus Simbol, Angka dan Emotikon

Ulasan yang ditulis pasti mengandung simbol dan angka. Selain itu, terdapat beberapa ulasan yang memuat emotikon sebagai salah satu bentuk penyampaian opini. Namun ketiganya tidak penting untuk diolah, sehingga perlu dihapus untuk memudahkan pengolahan data.

#### 3. Tokenisasi

Pada proses tokenisasi ini, tokenizer melakukan tugasnya untuk membagi sebuah kalimat menjadi beberapa bagian seperti kata-kata, frasa atau elemen bermakna yang lainnya [9].

#### 4. *Lemmatization*

Pada proses *lemmatization*, dilakukan transformasi untuk menormalisasi suatu kata dan mengubah kata menjadi bentuk dasarnya [9]. Contohnya adalah seperti yang ada pada data ulasan di penelitian ini, yaitu mengubah kata “tf” menjadi transfer, “tdk” menjadi tidak dan sebagainya.

## 5. Penghapusan *Stopword*

*Stopword* adalah kata umum yang sering muncul tetapi tidak memiliki pengaruh yang signifikan atau tidak memiliki makna [9]. Maka dari itu penghapusan *stopword* ini diperlukan. Namun sebelum dilakukannya proses ini peneliti harus membuat daftar kata *stopword* berdasarkan pada dataset yang ada. Contoh kata pada *stopword* adalah “pagi”, “misal”, “atau” dan lain sebagainya.

### 2.4 *Vector Space Model*

Data yang digunakan pada penelitian adalah berupa data teks yang mana termasuk pada data tidak terstruktur. Sedangkan komputer hanya mampu mengolah data terstruktur yang biasanya berbentuk tabular. Maka dari itu agar data ini bisa diolah oleh komputer, data perlu dikonversi menjadi angka. Sehingga disinilah peran dari VSM diperlukan untuk mengubah dokumen menjadi nilai vector.

Pada tahap ini, cara yang penulis gunakan adalah *tf-idf* (*term frequency and inverse document frequency*) bertujuan untuk memberi bobot pada kata  $t$  dalam dokumen  $d$  sesuai dengan rumus berikut:

$$(2.1) \quad weight(t, d) = tf(t, d) \times idf(t, D)$$

Dimana definisi dari  $t$ ,  $d$ ,  $D$ ,  $tf(t, d)$ ,  $idf(t, D)$  berturut-turut adalah kata, dokumen, corpus (kumpulan dokumen), *frequency*  $t$  di  $d$ , dan *inverse document frequency* dari  $t$  di  $D$ . Nilai *tf-idf* yang tertinggi adalah saat suatu kata  $t$  muncul berkali-kali dalam jumlah dokumen yang sedikit sedangkan nilai *tf-idf* menjadi lebih rendah apabila suatu kata  $t$  muncul lebih sedikit dalam satu dokumen, atau dalam banyak dokumen. Nilai *tf-idf* yang terendah adalah ketika kata muncul hampir di semua dokumen [10].

*Term frequency* (*tf*) akan menunjukkan seberapa banyak kata yang muncul dalam setiap dokumen. Hal ini menunjukkan tentang seberapa penting kata tersebut

dalam suatu dokumen. Semakin tingginya bobot tf menunjukkan bahwa semakin banyak kemunculan suatu kata dalam dokumen. Rumus tf adalah sebagai berikut:

$$tf = \begin{cases} 1 + \log_{10}(f_{t,d}), & f_{t,d} > 0 \\ 0, & f_{t,d} = 0 \end{cases} \quad (2.2)$$

*Inverse document frequency* (idf) menunjukkan tentang jarangnya suatu kata muncul. Kata yang jarang muncul berfungsi untuk membedakan satu dokumen dengan yang lainnya. Perhitungan dari idf adalah kebalikan dari df [11]. Rumus idf adalah sebagai berikut:

$$idf = \log_{10} \left( \frac{N}{df_t} \right) \quad (2.3)$$

Dimana  $N$  menunjukkan jumlah dari dokumen,  $df_t$  menunjukkan jumlah dari dokumen dalam corpus yang memuat kata  $t$ . Nilai idf yang tinggi menunjukkan jarangness kata tersebut muncul, sedangkan nilai idf yang rendah menunjukkan kata tersebut sering muncul [10].

## 2.5 Norm dan Dot Product

Panjang dari sebuah vektor  $\mathbf{v} = [v_1, v_2, \dots, v_n]$  pada  $\mathbb{R}^n$ , atau disebut juga sebagai *norm* didefinisikan sebagai berikut [12]:

$$\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + \dots + v_n^2}.$$

Jika  $\mathbf{u} = [u_1, u_2, \dots, u_n]$  dan  $\mathbf{v} = [v_1, v_2, \dots, v_n]$  merupakan vektor di  $\mathbb{R}^n$ , maka *dot product* didefinisikan sebagai berikut:

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + \dots + u_n v_n.$$

Vektor  $\mathbf{u} = (u_1, u_2, \dots, u_n)$  dan  $\mathbf{v} = (v_1, v_2, \dots, v_n)$  pada  $\mathbb{R}^n$  dapat direpresentasikan ke dalam matriks berukuran  $n \times 1$  sebagai berikut:

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \text{ dan } \mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}.$$

*Dot product* dari vektor  $u$  dan  $v$  dapat direpresentasikan sebagai perkalian matriks transpose  $u$  dengan matriks  $v$  sebagai berikut:

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T \mathbf{v} = [u_1 \quad u_2 \quad \cdots \quad u_n] \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = [u_1 v_1 \quad u_2 v_2 \quad \cdots \quad u_n v_n].$$

Sifat-sifat *dot product* adalah sebagai berikut [12]:

Jika  $\mathbf{u}, \mathbf{v}$  dan  $\mathbf{w}$  adalah vektor-vektor pada ruang berdimensi 2 atau berdimensi 3 dan  $k$  adalah skalar, maka:

1.  $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}.$
2.  $\mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}.$
3.  $k(\mathbf{u} \cdot \mathbf{v}) = (k\mathbf{u}) \cdot \mathbf{v} = \mathbf{u} \cdot (k\mathbf{v}).$
4.  $\mathbf{v} \cdot \mathbf{v} > 0$  jika  $\mathbf{v} \neq 0$ , dan  $\mathbf{v} \cdot \mathbf{v} = 0$  jika  $\mathbf{v} = 0$ .

## 2.6 Kernel

Kernel pada SVM digunakan untuk mentransformasi data ke ruang dengan dimensi yang lebih tinggi yang disebut sebagai ruang kernel [13]. Berikut ini akan ditunjukkan contoh ilustrasi dari pemisahan data dengan menggunakan kernel. Misal dua buah data dinyatakan sebagai  $x_i = (u_1, z_1)$  dan  $x_j = (u_2, z_2)$ . Diasumsikan fungsi kernel akan dibuat dengan menggunakan kedua data tersebut.

$$K(x_i, x_j) = (x_i \cdot x_j^T)^2$$

$$K(x_i, x_j) = (u_i u_2 + z_1 z_2)^2$$

$$K(x_i, x_j) = (u_1^2 u_2^2 + z_1^2 z_2^2 + 2 u_1 u_2 z_1 z_2)$$

$$K(x_i, x_j) = (u_1, \sqrt{2} u_1 z_1, z_1) (u_2, \sqrt{2} u_2 z_2, z_2)$$

$$K(x_i, x_j) = \phi(x_i) \phi(x_j)^T$$

Nilai  $K$  yang telah disebutkan di atas telah mendefinisikan pemetaan ke ruang dengan dimensi yang lebih tinggi seperti berikut:

$$\phi(x_i) = \{u_1, \sqrt{2} u_1 z_1, z_1\}$$

Contoh numerik dari kernel adalah sebagai berikut. Misal  $x_i = (5,3)$  dan  $x_j = (2,5)$ , maka:

$$\begin{aligned}K(x_i, x_j) &= (x_i \cdot x_j^T)^2 \\&= (5 \cdot 2 + 3 \cdot 5)^2 \\&= (10 + 15)^2 \\&= 25^2\end{aligned}$$

$$K(x_i, x_j) = 625$$

Berikut ini merupakan tiga fungsi kernel yang diujikan pada penelitian ini:

#### 1. Kernel Linier

Kernel linear merupakan fungsi kernel yang paling sederhana. Kernel ini cocok ketika terdapat banyak fitur dikarenakan pemetaan ke ruang dimensi yang lebih tinggi tidak benar-benar meningkatkan kinerja, contohnya adalah pada klasifikasi teks. Dalam klasifikasi teks, baik jumlah dokumen maupun jumlah fitur (kata) sama-sama besar. Berikut merupakan persamaan dari kernel linear:

$$K(x_i, x_j) = (x_i \cdot x_j^T)$$

#### 2. Kernel *Radial Basis Function*

Kernel *Radial Basis Function* (RBF) merupakan fungsi kernel yang digunakan ketika data tidak dapat dipisahkan secara linear. RBF kernel memiliki dua parameter yaitu Gamma dan Cost. Tugas dari parameter Cost atau C ini adalah sebagai pengoptimalan SVM untuk menghindari terjadinya misklasifikasi di setiap sampel dalam training dataset. Sedangkan tugas dari parameter Gamma adalah untuk menentukan pengaruh dari satu sampel training dataset pada garis pemisahnya. Berikut merupakan persamaan dari kernel RBF:

$$K(x_i, x_j) = \exp[-\gamma \|x - z\|^2]$$



### 3. Kernel Polynomial

Kernel polinomial merupakan fungsi kernel yang digunakan ketika data tidak dapat dipisahkan secara linear. Kernel ini cocok digunakan untuk permasalahan dimana semua training dataset dinormalisasi. Berikut merupakan persamaan dari kernel polynomial:

$$K(x_i, x_j) = (x_i \cdot x_j^T)^d$$

#### 2.7 Grid Search Cross Validation

*Grid Search Cross Validation* (*Grid Search CV*) merupakan salah satu proses untuk melakukan pemilihan *hyperparameter* terbaik untuk suatu model yang diberikan. *Hyperparameter* adalah parameter yang ditentukan tanpa proses uji atau dengan kata lain merupakan parameter yang tidak ditentukan oleh mesin. *Hyperparameter* yang diperoleh merupakan parameter terbaik yang akan dimasukkan pada model [14]. *Grid Search CV* melakukan kombinasi dari *hyperparameter* yang telah ditentukan dan menghitung rata-rata nilai *cross validation* dari setiap kombinasinya. Kombinasi dengan rata-rata nilai *cross validation* yang akan dimasukkan pada model.

Contoh sederhana untuk menggambarkan cara kerja dari *Grid Search CV* ini adalah misalkan *hyperparameter*  $X = [2,5]$  dan  $Y = [0,3]$ . Kemudian *Grid Search CV* melakukan kombinasi  $X$  dan  $Y$  dengan hasil  $[2,0]$ ,  $[2,3]$ ,  $[5,0]$  dan  $[5,3]$ . Setelah itu, akan dipilih kombinasi terbaik berdasarkan rata-rata nilai *cross validation* tertinggi.

#### 2.8 Evaluasi Model

Evaluasi model perlu dilakukan untuk mengetahui seberapa baik suatu model dapat mengklasifikasikan suatu kelas. Salah satu cara untuk melakukan hal tersebut adalah dengan menggunakan *confusion matrix*. *Confusion matrix* adalah sebuah tabel yang menyatakan berapa banyak data uji yang benar dan salah

diklasifikasikan. Parameter yang digunakan pada data uji yaitu TP (*true positive*), FN (*false negative*), TN (*true negative*), dan FP (*false positive*) [15].

**Tabel 2. 1** *Confusion Matrix*

Aktual	Prediksi	
	Negatif	Positif
Negatif	TN	FN
Positif	FP	TP

Dari *confusion matrix* tersebut dapat dihasilkan nilai akurasi [16]. Nilai akurasi digunakan untuk mengukur seberapa akurat suatu model dalam mengklasifikasikan suatu kelas dengan benar. Formula untuk menghitung nilai akurasi adalah:

$$\text{Akurasi} = \frac{TN+TP}{TN+FP+TP+FN} \quad (2.1)$$

Contoh kasus yang sederhana adalah misalkan Dinas Lingkungan Hidup dan Kebersihan ingin mengukur kinerja dari sebuah mesin pemisah yang bertugas memisahkan sampah organik dari semua sampah yang ada. Untuk mengujinya, petugas memasukkan 100 sampah organik dan 900 sampah lain. Hasilnya mesin tersebut memisahkan 110 yang dideteksi sebagai sampah organik. Kemudian setelah dicek kembali oleh petugas, ternyata dari 110 sampah tersebut hanya 90 barang yang merupakan sampah organik, sedangkan 20 lainnya merupakan sampah lain. Sehingga perhitungan dari akurasinya adalah sebagai berikut:

$$\text{Akurasi} = \frac{90+880}{90+20+10+880} = \frac{970}{1000} = 0,97$$

## BAB III

### METODELOGI PENELITIAN

Bab ini menjelaskan metode-metode yang digunakan dalam penelitian secara teori dan contoh penerapannya. Metode yang digunakan antara lain SVM sebagai metode klasifikasi sentiment dan aspek. Pada bab ini juga akan dijelaskan bagaimana alur penelitian ABSA.

#### 3.1 Sumber Data

Data yang digunakan pada penelitian ini merupakan data sekunder berupa data ulasan atau *review* berbahasa Indonesia yang diberikan oleh pengguna Tokopedia melalui situs *Google Play Store*. Data yang diperoleh dengan menggunakan teknik *scraping* ini berjumlah 5.614 ulasan dari bulan April sampai dengan Juli 2019.

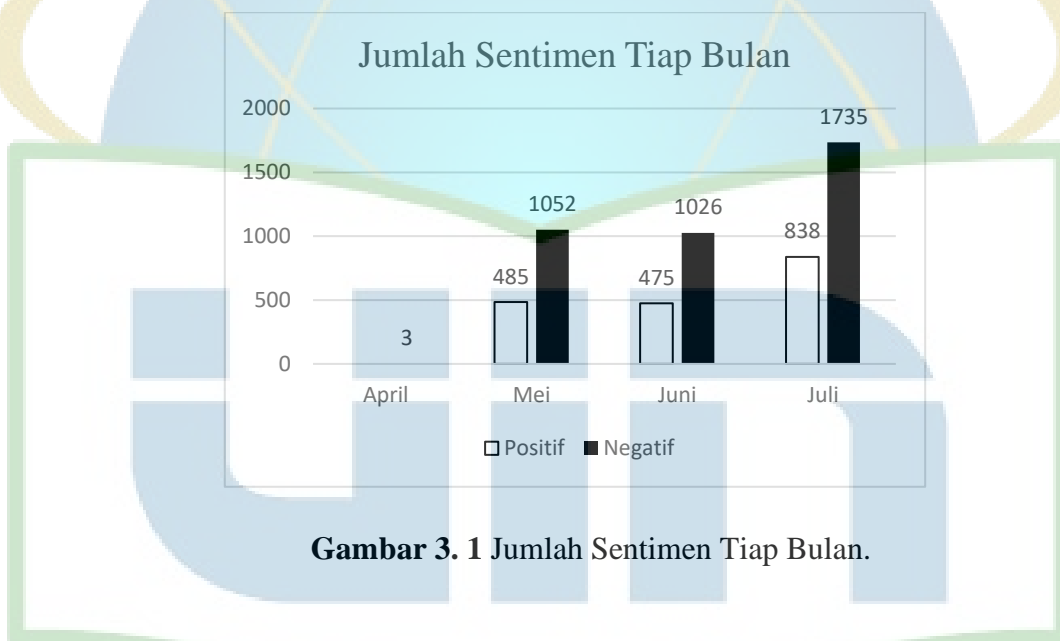
Pada saat proses *scraping*, peneliti menggunakan *package Selenium* dan browser Chrome. Ulasan yang didapat dari hasil *scraping* tidakurut waktu, melainkan terurut berdasarkan yang paling membantu (*most helpful*). Data berisikan dua kolom, diantaranya adalah *date* yang merupakan tanggal pada saat pengguna memberikan ulasannya dan *review* yang berisikan kalimat ulasan dari pengguna tentang aplikasi Tokopedia. Data disimpan dalam bentuk *Comma Separated Values* (CSV). Berikut adalah beberapa data awal dari hasil *scraping*:

**Tabel 3. 1** Data Awal Ulasan Tokopedia.

DATE	REVIEW
20 Juli 2019	Baru kali ini belanja di tokped, barang yang saya beli di bawa kabur ekspedisi, aduuh kecewa jadinya, proses claim asuransi sudah mau 2 minggu <b>dan itu</b> pun tidak jelas sampai kapan, dan tidak ada pemberitahuan yang jelas di forum komplain! bisa tolong di bantu proses alur untuk pengembalian dana berapa lama prosesnya sampai masuk ke saldo? saya sebagai pembeli jadi kurang nyaman dan waswas belanja disini. terimakasih..
19 Juli 2019	Toko Pedia Adalah aplikasi belanja online, dimana pada aplikasi ini kalian bisa berbelanja berbagai macam barang ada hanphone, baju, mainan anak dan masih banyak lagi, kelebihan dari aplikasi ini menurut saya cukup mudah untuk di gunakan, selain itu ada banyak sekali fitur yang ada pada aplikasi ini, barang-barang yang di jual juga cukup banyak, bahkan menurut saya sangat lengkap. Dan pada aplikasi ini selain menjual barang-barang elektronik <b>dan rumah tangga ternyata aplikasi ini bisa memesan tr</b>
13 Juli 2019	aplikasinya <b>kok</b> gak bisa di akses yaa knp dgn aplikasinya, saya mendapat pemberitahuan bahwasannya saya berhak mendapatkan hadiah smartphone samsung, untuk mendapatkannya, terlebih dahulu saya harus menginstal apknya terlebih dahulu untuk bisa mendapatkan hadiah tersebut, jadi saya mohon untuk di perbaiki kesalahan ini terima kasih.
20 Juli 2019	Tolong dibikin menu untuk melaporkan pedagang/item barang yg upload iklan tapi stok barang kosong,. loading gambar sangat lambat bahkan tidak keluar gambarnya setelah ditunggu lama,. udah ganti perangkat hp juga sama sering gak muncul gambar,. perangkat menggunakan samsung j2, xiaomi 4prime, moto 4plus, <b>provider indosat, axis bahkan hotspot spedy juga</b> gak bisa keluar gambar
14 Juli 2019	Tolong aplikasi Tokopedia setelah update terbaru kenapa gambarnya malah tidak muncul hampir semua di produknya,foto review,bahkan foto profil saya juga tidak muncul tolong perbaiki,,,padahal sebelumnya semuanya normal2 saja,saya sering belanja di tokopedia kalau mau beli lihat gambarnya gak bisa terus gimana mau memilih barang.terimakasih

### 3.2 Pelabelan Sentimen

Setelah data diperoleh melalui *web scraping*, kemudian dilakukan analisis dengan memberikan label pada setiap ulasan. Kelas sentimen yang digunakan yaitu positif dan negatif dan label yang digunakan yaitu 1 untuk ulasan bernilai positif dan -1 untuk ulasan bernilai negatif. Suatu ulasan dikatakan bernilai positif apabila ulasan berisikan tentang pengalaman berbelanja yang menyenangkan dan memuaskan dan bernilai negatif apabila berisikan tentang keluhan dan kekecewaan dari pengguna. Pelabelan dilakukan dengan tujuan untuk memberikan pembelajaran pada model.

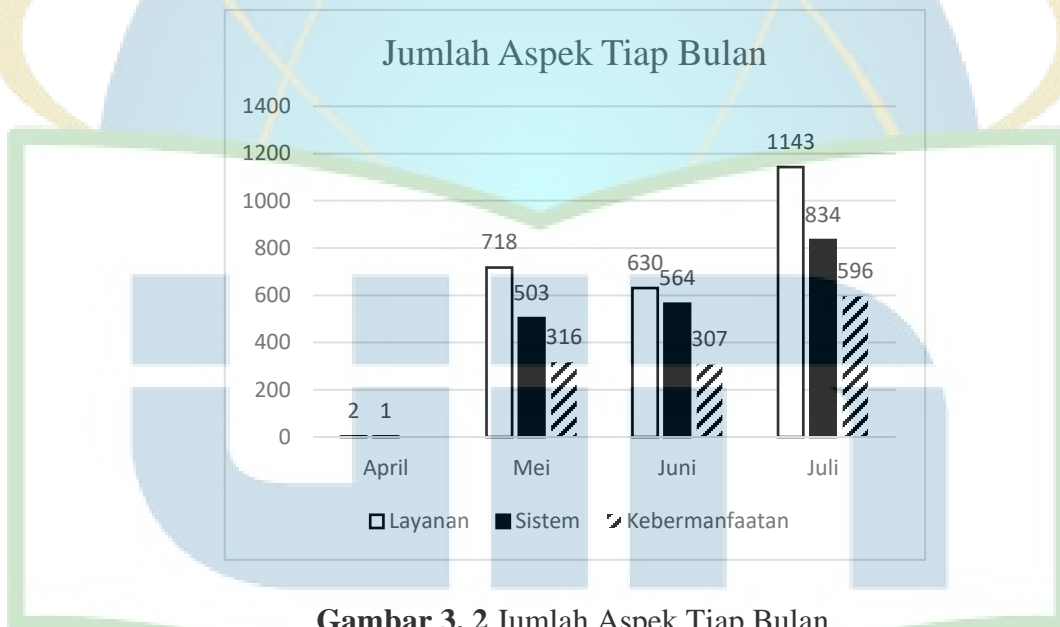


**Gambar 3. 1** Jumlah Sentimen Tiap Bulan.

Hasil pelabelan yang dilakukan adalah seperti pada Gambar 3.2, dapat diketahui bahwa pada setiap bulannya jumlah ulasan yang bersentimen negatif lebih banyak dari yang bersentimen positif. Selain itu, jumlah ulasan yang bersentimen negatif paling banyak berada pada bulan Juli yaitu sebanyak 1.735 ulasan.

### 3.3 Pelabelan Aspek

Setelah melakukan pelabelan sentiment, tahap selanjutnya adalah melakukan pelabelan aspek. Aspek yang digunakan adalah layanan, sistem dan kebermanfaatan. Suatu ulasan termasuk dalam aspek layanan apabila memuat tentang kecepatan respon saat terjadi kendala. Sedangkan yang termasuk pada aspek sistem adalah ulasan yang berisi tentang performa dari aplikasi Tokopedia saat digunakan. Dan termasuk pada kategori kebermanfaatan apabila pada ulasan terdapat kata-kata yang menunjukkan bahwa pengguna merasa terbantu dengan adanya Tokopedia. Ketiga aspek ini ditentukan dengan cara penulis membaca data ulasan sehingga dapat diketahui aspek apa saja yang paling banyak dibahas pada ulasan.



**Gambar 3. 2** Jumlah Aspek Tiap Bulan.

Jumlah dari pelabelan aspek setiap bulannya yang telah dilakukan dapat dilihat pada Gambar 3.3. Aspek layanan selalu terlihat lebih banyak dari aspek lainnya, yaitu sebanyak 2 ulasan pada bulan April, 718 ulasan pada bulan Mei, 625 ulasan pada bulan Juni dan 1.143 ulasan pada bulan Juli.

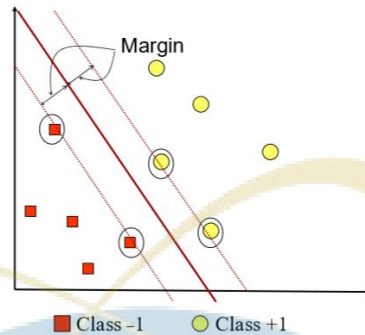
### 3.4 Data Training dan Data Testing

Tahap lanjutan yang dilakukan setelah data diubah menjadi vector adalah membagi data menjadi data training, validasi dan testing. Data training atau data yang dilatih untuk membangun model dibentuk dengan ukuran 80% dari data yaitu sebanyak 4.491 data dan untuk 20% dari sisanya atau sebanyak 1.123 data digunakan sebagai data testing untuk menguji performa model yang sudah dilatih.

### 3.5 Support Vector Machine

Metode klasifikasi yang digunakan pada penelitian ini adalah SVM. Konsep dasar dari metode yang diperkenalkan oleh Vladimir Vapnik, Boser dan Guyon pada tahun 1992 ini adalah mentransformasi data ke ruang yang berdimensi lebih tinggi dan menemukan *hyperplane* terbaik [17]. Hyperplane adalah bidang datar penentu yang memisahkan dua buah kelas di dimensi  $n$ . Untuk menemukan hyperplane terbaik adalah dengan cara mengukur *margin* hyperplane tersebut. Margin adalah jarak antara hyperplane dengan pattern terdekat dari masing-masing kelas. Pattern yang paling dekat dengan hyperplane disebut *support vector* [18].

Misalkan data latih dinyatakan sebagai  $(\mathbf{x}_i, y_i)$  dimana  $i = 1, 2, \dots, n$ .  $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{ij}]$  adalah vektor baris dari fitur ke-  $i$  di ruang dimensi ke-  $j$  dan  $y_i$  adalah label dari  $\mathbf{x}_i$  yang didefinisikan sebagai  $y_i \in \{+1, -1\}$ . Diasumsikan kedua kelas -1 dan +1 dapat dipisah secara linear oleh hyperplane. Pada gambar 3.1 hyperplane ditunjukkan dengan garis lurus berwarna merah. Data yang berada di atas hyperplane adalah kelas +1 dan data yang berada di bawah hyperplane adalah kelas -1.



**Gambar 3. 3** Contoh *Hyperplane* Dua Dimensi [15].

Persamaan hyperplane didefinisikan sebagai berikut:

$$f(x) = \mathbf{w} \cdot \mathbf{x} + b, \quad (3.1)$$

dengan:

$\mathbf{w}$  = parameter bobot,

$\mathbf{x}$  = vektor input,

$b$  = bias.

Vektor  $\mathbf{w}$  memiliki arah tegak lurus dengan hyperplane. Jika nilai  $b$  berubah maka hyperplane akan berubah juga. Hyperplane terbaik adalah hyperplane yang terletak di tengah-tengah antara dua set obyek dari dua kelas. Untuk itu, perlu menemukan hyperplane terbaik dengan mendapatkan nilai margin terbesar. Margin terbesar dapat ditemukan dengan memaksimalkan nilai jarak antara hyperplane dan titik terdekatnya. Pattern yang memenuhi kelas -1 adalah pattern yang memenuhi persamaan  $\mathbf{w} \cdot \mathbf{x}_i + b = -1$  dan pattern yang memenuhi kelas +1 adalah pattern yang memenuhi persamaan  $\mathbf{w} \cdot \mathbf{x}_i + b = 1$ .

Support vektor direpresentasikan sebagai titik  $(x, y)$ . Hyperplane sebagai berikut:

$$Ax + By + C = 0, \quad (3.2)$$

dengan rumus jarak sebagai berikut:

$$d = \frac{|Ax + By + C|}{\sqrt{A^2 + B^2}}.$$



Persamaan (3.2) diubah dalam bentuk *dot product* pada vektor sehingga menjadi:

$$[A \ B] \begin{bmatrix} x \\ y \end{bmatrix} + C = 0.$$

Misalkan  $\mathbf{w} = [A \ B]$  dan  $\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$  dan  $b = C$ , maka diperoleh:

$$d = \frac{|Ax + By + C|}{\sqrt{A^2 + B^2}} = \frac{|\mathbf{w} \cdot \mathbf{x} + b|}{\sqrt{\mathbf{w}^2 + C^2}} = \frac{|\mathbf{w} \cdot \mathbf{x} + b|}{\sqrt{\mathbf{w}^2}} = \frac{|\mathbf{w} \cdot \mathbf{x} + b|}{\|\mathbf{w}\|}.$$

Nilai margin dapat dicari menggunakan nilai tengah antara jarak kedua kelas sebagai berikut:

$$\begin{aligned} \text{margin} &= \frac{1}{2} (d^+ - d^-) \\ &= \frac{1}{2} \left( \frac{|\mathbf{w} \cdot \mathbf{x}_1 + b|}{\|\mathbf{w}\|} - \frac{|\mathbf{w} \cdot \mathbf{x}_2 + b|}{\|\mathbf{w}\|} \right) \\ &= \frac{1}{2} \left( \frac{1}{\|\mathbf{w}\|} - \frac{(-1)}{\|\mathbf{w}\|} \right) \\ &= \frac{1}{\|\mathbf{w}\|}, \|\mathbf{w}\| \neq 0, \end{aligned}$$

dimana:

$d^+$  : jarak antara hyperplane terhadap kelas +1,

$d^-$  : jarak antara hyperplane terhadap kelas -1.

Setiap kelas harus ditambahkan batasan pada data dari masing-masing kelas agar tidak masuk ke dalam margin, batasannya sebagai berikut:

$$\mathbf{w} \cdot \mathbf{x}_i + b \leq -1, \text{ jika } y = -1,$$

$$\mathbf{w} \cdot \mathbf{x}_i + b \geq +1, \text{ jika } y = +1,$$

atau dapat ditulis sebagai berikut:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0, \forall 1 \leq i \leq n, i \in N.$$

Memaksimalkan nilai margin ekuivalen dengan meminimumkan  $\|\mathbf{w}\|^2$ . Maka pencarian hyperplane terbaik dengan nilai margin terbesar dapat dirumuskan menjadi masalah optimasi pemrograman kuadratik sebagai berikut:

$$\max \text{margin} = \min \frac{1}{2} \|\mathbf{w}\|^2,$$

dengan kendala:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0, \forall 1 \leq i \leq n, i \in N.$$

Masalah ini dapat diselesaikan dengan mengubah persamaan ke dalam fungsi lagrange:

$$\min L_p(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1],$$

dimana:

$L_p$ : fungsi lagrange (*primal problem*),

$\alpha_i$ : nilai dari koefisien lagrange,  $\alpha_i \geq 0$  dengan  $i = 1, 2, \dots, n$ .

Fungsi  $L_p$  diminimumkan terhadap  $\mathbf{w}$  dan  $b$  dan dimaksimumkan terhadap  $\alpha$ , sehingga akan dicari turunan pertama dari fungsi  $L_p$  terhadap  $\mathbf{w}$  dan  $b$ , maka didapat:

1. Turunan pertama fungsi  $L_p$  terhadap  $\mathbf{w}$

$$\frac{\partial}{\partial \mathbf{w}} L_p(\mathbf{w}, b, \alpha) = 0.$$

Maka akan didapatkan:

$$\min L_p(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b)] + \sum_{i=1}^n \alpha_i,$$

$$\frac{\partial}{\partial \mathbf{w}} L_p(\mathbf{w}, b, \alpha) = \mathbf{w} - \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i$$

$$\Leftrightarrow 0 = \mathbf{w} - \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i$$

$$\Leftrightarrow \mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i. \quad (3.3)$$

2. Turunan pertama fungsi  $L_p$  terhadap  $b$

$$\frac{\partial}{\partial b} L_p(\mathbf{w}, b, \alpha) = 0.$$

Maka akan didapatkan:

$$\min L_p(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i + b)] + \sum_{i=1}^n \alpha_i,$$

$$\frac{\partial}{\partial b} L_p(\mathbf{w}, b, \alpha) = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i$$

$$\Leftrightarrow 0 = \sum_{i=1}^n \alpha_i y_i \cdot$$

Formula langrange  $L_p$  (primal problem) diubah menjadi  $L_D$  (dual problem).

$$\begin{aligned} maks L_D(\alpha) &= \frac{1}{2} \left( \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \right) \left( \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \right) - \sum_{i=1}^n \alpha_i y_i \left( \left( \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \right) \mathbf{x}_i + b \right) \\ &\quad + \alpha_i \\ &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i y_i \alpha_j y_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^n \sum_{j=1}^n \alpha_i y_i \alpha_j y_j (\mathbf{x}_i \cdot \mathbf{x}_j) - b \\ &= \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i y_i \alpha_j y_j (\mathbf{x}_i \cdot \mathbf{x}_j), \end{aligned} \quad (3.4)$$

dengan kendala,

$$\sum_{i=1}^n \alpha_i y_i = 0, \alpha_i \geq 0.$$

Nilai  $\alpha_i$  diperoleh dari hasil perhitungan substitusi kendala pada persamaan (3.4).

Nilai  $\alpha_i$  akan digunakan untuk menemukan nilai  $\mathbf{w}$ . Setiap titik data selalu terjadi  $\alpha_i = 0$ . Titik-titik data dimana  $\alpha_i = 0$  tidak akan muncul dalam perhitungan mencari nilai  $\mathbf{w}$  sehingga tidak berperan dalam memprediksi data baru. Data lain dimana  $\alpha_i > 0$  disebut support vector.

Dilakukan  $\text{sign}\{f(x)\}$  untuk menguji data baru menggunakan model yang sudah dilatih. Substitusikan persamaan (3.3) ke persamaan (3.1) dan menggunakan kernel linear  $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x} \cdot \mathbf{x}^T$  sehingga diperoleh:

$$f(x) = \sum_{i=1}^n \alpha_i y_i (\mathbf{x}_i^T \cdot \mathbf{x}) + b. \quad (3.5)$$

Mensubstitusikan persamaan (3.5) ke dalam  $y_i f(\mathbf{x}_i) = 1$  diperoleh:

$$y_i \sum_{m \in S} \alpha_m y_m \mathbf{x}_m^T \cdot \mathbf{x}_i + b = 1,$$

dimana  $S$  adalah himpunan indeks support vector.

Nilai  $b$  diperoleh sebagai berikut:

$$\begin{aligned} y_i \left( \sum_{m \in S} \alpha_m y_m \mathbf{x}_m^T \cdot \mathbf{x}_i + b \right) &= 1 \\ \Leftrightarrow y_i y_i \left( \sum_{m \in S} \alpha_m y_m \mathbf{x}_m^T \cdot \mathbf{x}_i + b \right) &= y_i \\ \Leftrightarrow \left( \sum_{m \in S} \alpha_m y_m \mathbf{x}_m^T \cdot \mathbf{x}_i + b \right) &= y_i \\ \Leftrightarrow b &= y_i - \sum_{m \in S} \alpha_m y_m \mathbf{x}_m^T \cdot \mathbf{x}_i \\ \Leftrightarrow b &= \frac{1}{N_S} \sum_{i \in S} \left( y_i - \sum_{m \in S} \alpha_m y_m \mathbf{x}_m^T \cdot \mathbf{x}_i \right) \end{aligned} \quad (3.6)$$

dimana  $N_S$  adalah jumlah support vector.

### 3.6 Analisa Numerik SVM

Diberikan contoh data penerapan metode SVM linear:

Misalkan terdapat data (1,2) pada kelas -1 dan data (3,4) pada kelas +1. Dengan menggunakan kedua kelas tersebut dibuat model yang memprediksi kelas (0,1).

Menggunakan persamaan (3.4) diperoleh:

$$\begin{aligned}
 L_D(\alpha) &= \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i y_i \alpha_j y_j (x_i \cdot x_j) \\
 &= \sum_{i=1}^2 \alpha_i - \frac{1}{2} \sum_{i=1}^2 \sum_{j=1}^2 \alpha_i y_i \alpha_j y_j (x_i \cdot x_j) \\
 &= \alpha_1 + \alpha_2 - \frac{1}{2} (\alpha_1 \alpha_1 y_1 y_1 (x_1 \cdot x_1) + \alpha_1 \alpha_2 y_1 y_2 (x_1 \cdot x_2) \\
 &\quad + \alpha_2 \alpha_1 y_2 y_1 (x_2 \cdot x_1) + \alpha_2 \alpha_2 y_2 y_2 (x_2 \cdot x_2)) \\
 &= \alpha_1 + \alpha_2 - \frac{1}{2} (\alpha_1^2 (-1)(-1) \binom{1}{2} \cdot \binom{1}{2} + \alpha_1 \alpha_2 (-1)(1) \binom{1}{2} \cdot \binom{3}{4} \\
 &\quad + \alpha_2 \alpha_1 (1)(-1) \binom{3}{4} \cdot \binom{1}{2} + \alpha_2^2 (1)(1) \binom{3}{4} \binom{3}{4}) \\
 &= \alpha_1 + \alpha_2 - \frac{1}{2} (5\alpha_1^2 - 11\alpha_1 \alpha_2 - 11\alpha_2 \alpha_1 + 25\alpha_2^2) \\
 &= \alpha_1 + \alpha_2 - \frac{5}{2} \alpha_1^2 + 11\alpha_1 \alpha_2 - \frac{25}{2} \alpha_2^2, \\
 &\text{dengan } \sum_{i=1}^2 \alpha_i y_i = 0 \\
 &\Leftrightarrow \alpha_1 y_1 + \alpha_2 y_2 = 0 \\
 &\Leftrightarrow \alpha_1 (-1) + \alpha_2 (1) = 0 \\
 &\Leftrightarrow \alpha_1 = \alpha_2.
 \end{aligned}$$

Substitusi  $\alpha_1 = \alpha_2$  ke persamaan  $L_D(\alpha)$  sehingga:

$$\begin{aligned} L_D(\alpha) &= \alpha_1 + \alpha_1 - \frac{5}{2}\alpha_1^2 + 11\alpha_1\alpha_1 - \frac{25}{2}\alpha_1^2 \\ &= 2\alpha_1 - \frac{5}{2}\alpha_1^2 + 11\alpha_1^2 - \frac{25}{2}\alpha_1^2 \\ &= 2\alpha_1 - 4\alpha_1^2. \end{aligned}$$

Lalu  $L_D(\alpha)$  diturunkan terhadap  $\alpha_1$  diperoleh:

$$\begin{aligned} \frac{\partial L}{\partial \alpha_1} L_D(\alpha) &= 0 \\ \Leftrightarrow \frac{\partial L}{\partial \alpha_1} (2\alpha_1 - 4\alpha_1^2) &= 0 \\ \Leftrightarrow 2 - 8\alpha_1 &= 0 \\ \Leftrightarrow \alpha_1 &= \frac{1}{4}. \end{aligned}$$

Karena  $\alpha_1 = \alpha_2$  dan  $\alpha_1 = \frac{1}{4}$  maka  $\alpha_2 = \frac{1}{4}$  sehingga menggunakan persamaan (3.3) diperoleh nilai  $w$  sebagai berikut:

$$w = \sum_{i=1}^2 \alpha_i y_i x_i = \alpha_1 y_1 x_1 + \alpha_2 y_2 x_2 = \frac{1}{4}(-1)\begin{pmatrix} 1 \\ 2 \end{pmatrix} + \frac{1}{4}(1)\begin{pmatrix} 3 \\ 4 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}.$$

Mencari nilai  $b$  menggunakan persamaan (3.6) sehingga diperoleh:

$$\begin{aligned} b &= \frac{1}{2} \sum_{j=1}^2 \left( y_j - \sum_{i=1}^2 \alpha_i y_i (x_i^T \cdot x_j) \right) \\ &= \frac{1}{2} \sum_{j=1}^2 \left( y_j - (\alpha_1 y_1 (x_1^T \cdot x_j) + \alpha_2 y_2 (x_2^T \cdot x_j)) \right) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2} \left( y_1 - \left( \alpha_1 y_1 (x_1^T \cdot x_1) + \alpha_2 y_2 (x_2^T \cdot x_1) \right) + y_2 - \left( \alpha_1 y_1 (x_1^T \cdot x_2) + \alpha_2 y_2 (x_2^T \cdot x_2) \right) \right) \\
&= \frac{1}{2} \left( (-1) - \left( \frac{1}{4} (-1) (1 \ 2) \cdot \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \frac{1}{4} (1) (3 \ 4) \cdot \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right) + 1 \left( \frac{1}{4} (-1) (1 \ 2) \cdot \begin{pmatrix} 3 \\ 4 \end{pmatrix} + \frac{1}{4} (1) (3 \ 4) \cdot \begin{pmatrix} 3 \\ 4 \end{pmatrix} \right) \right) \\
&= \frac{10}{4},
\end{aligned}$$

sehingga  $f(x) = w \cdot x + b = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix} x + \frac{10}{4}$ ,

maka  $sign(f(x)) = sign\left(\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \frac{10}{4}\right) = sign(3) = +1$ ,

dengan mengevaluasi tanda dari  $f(x)$  diperoleh kelas dari (0,1) adalah +1.

### 3.7 Soft Margin

Prinsip kerja dari model SVM adalah menentukan *hyperplane* terbaik dari suatu *dataset* untuk dipisahkan berdasarkan kelasnya. Paada kenyataannya, data yang sering dijumpai adalah data yang tidak dapat dipisahkan secara linear, yaitu saat tidak ada sebuah garis atau bidang yang dapat memisahkan antar kelas pada data [19]. Maka untuk mengatasi hal tersebut dibutuhkan *soft margin classifier* yaitu dengan mengklasifikasikan sebagian besar data benar dan memberikan kemungkinan pada model untuk terjadi kesalahan saat mengklasifikasi pada beberapa titik di sekitar bidang pemisah [13]. Bidang pemisah harus diubah sehingga lebih fleksibel dengan cara menambahkan variabel *slack*  $\xi_i$  dimana  $\xi > 0$  sehingga menjadi  $x_i \cdot w + b \geq 1 - \xi$  untuk kelas 1 dan  $x_i \cdot w + b \leq -1 + \xi$  untuk kelas 2 [20]. Dengan adanya penambahan variabel  $\xi_i$  yang bertujuan untuk menunjukkan ketelitian pemisahan yang

memungkinkan suatu titik berada pada kondisi misklasifikasi, pencarian *hyperplane* terbaik dirumuskan sebagai berikut:

$$\max \text{margin} = \min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i$$

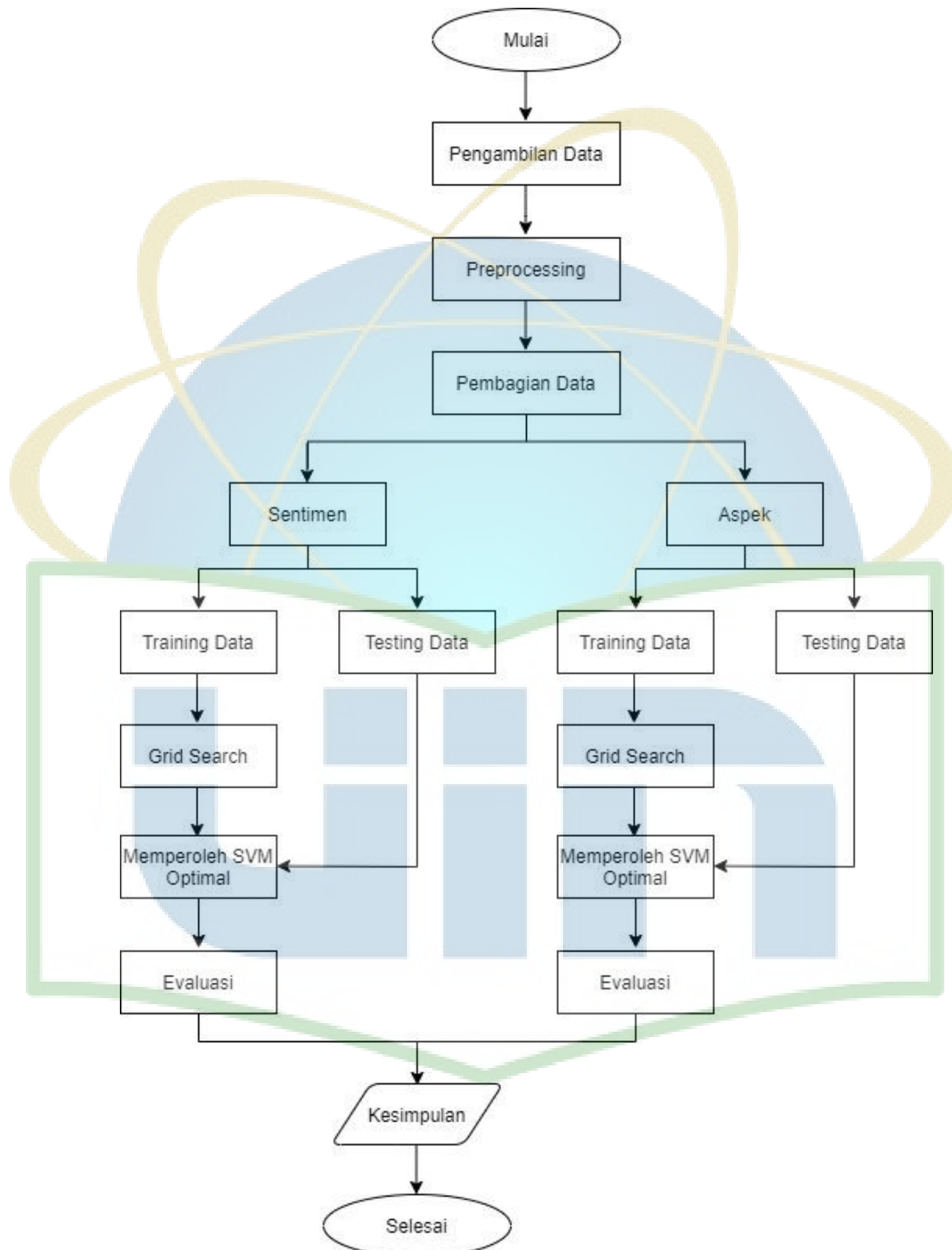
C adalah parameter untuk menghindari kesalahan pada saat klasifikasi [21]. Meminimumkan  $C \sum_{i=1}^n \xi_i$  adalah untuk meminimumkan *error* pada data latih. Semakin besar nilai C, semakin besar pula pelanggaran yang dikenakan pada tiap klasifikasi [22].

Perubahan juga terjadi pada fungsi lagrange primal, yaitu:

$$\min L_p(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^n \mu_i \xi_i$$



### 3.8 Alur Penelitian



**Gambar 3. 4** Alur Penelitian

## BAB IV

### HASIL DAN PEMBAHASAN

Pada bab ini penulis akan menjelaskan hasil dari pengambilan data, data yang telah dipreprocessing dan wordcloud dari setiap kelas berdasarkan labeling. Kemudian menjelaskan grafik hasil training data serta menunjukkan hasil uji coba beberapa model machine learning untuk menentukan model mana yang paling optimal. Lalu akan ditunjukkan pula evaluasi model dengan *confusion matrix*.

#### 4.1 Hasil Preprocessing dan Text Analytics

Data yang diperoleh melalui proses *scraping* perlu dibersihkan terlebih dahulu sebelum diolah oleh mesin. Proses ini dinamakan preprocessing. Namun sebelum preprocessing, terlebih dahulu dilakukan pelabelan aspek dan sentimen.

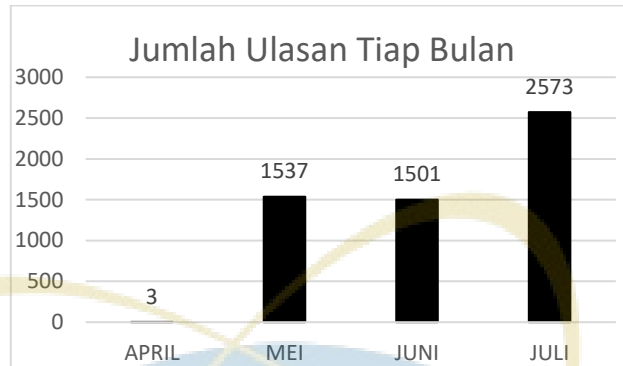
Adapun langkah preprocessing yang dilakukan diantaranya adalah case folding, menghapus simbol, angka dan emotikon, lemmatization, penghapusan stopwords serta tokenizer. Namun pada tahap penghapusan stopwords tidak menghapus kata negasi seperti “tidak” agar tidak salah dalam menginterpretasikan data. Hasil dari proses preprocessing dan pelabelan adalah sebagai berikut:

**Tabel 4. 1** Hasil *Preprocessing* dan Pelabelan

REVIEW	REVIEW BERSIH	LABEL	ASPECT
Baru kali ini belanja di tokped, barang yang saya beli di bawa kabur ekspedisi, aduuh kecewa jadinya, proses claim asuransi sudah mau 2 minggu dan itu pun tidak jelas sampai kapan, dan tidak ada pemberitahuan yang jelas di forum komplain! bisa tolong di bantu proses alur untuk pengembalian dana berapa lama prosesnya sampai masuk ke saldo? saya sebagai pembeli jadi kurang nyaman dan waswas belanja disini. terimakasih..	belanja barang beli bawa kabur ekspedisi kecewa proses claim asuransi minggu forum keluh tolong bantu proses alur dana prosesnya masuk saldo beli nyaman waswas belanja terimakasih	-1	SERVICE

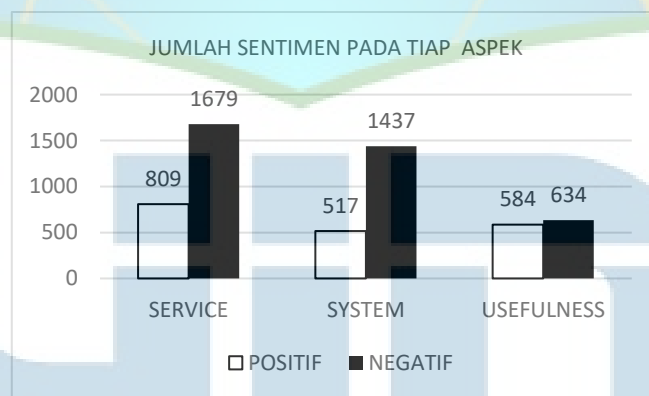
Toko Pedia Adalah aplikasi belanja online, dimana pada aplikasi ini kalian bisa berbelanja berbagai macam barang ada handphone, baju, mainan anak dan masih banyak lagi, kelebihan dari aplikasi ini menurut saya cukup mudah untuk di gunakan, selain itu ada banyak sekali fitur yang ada pada aplikasi ini, barang-barang yang di jual juga cukup banyak, bahkan menurut saya sangat lengkap. Dan pada aplikasi ini selain menjual barang-barang elektronik dan rumah tangga ternyata aplikasi ini bisa memesan tr	toko pedia aplikasi belanja online dimana aplikasi belanja barang handphone baju main anak aplikasi mudah fitur aplikasi barang barang jual lengkap aplikasi jual barang barang elektronik rumah tangga nyata aplikasi pesan	1	USEFULLNESS
aplikasinya kok gak bisa di akses yaa knp dgn aplikasinya, saya mendapat pemberitahuan bahwasannya saya berhak mendapatkan hadiah smartphone samsung, untuk mendapatkannya, terlebih dahulu saya harus menginstal apknya terlebih dahulu untuk bisa mendapatkan hadiah tersebut, jadi saya mohon untuk di perbaiki kesalahan ini terima kasih.	aplikasinya akses aplikasinya bahwasannya hak hadiah smartphone pasang aplikasinya hadiah mohon perbaiki salah terima kasih	-1	SYSTEM

Pada Tabel 4.1, hasil kalimat ulasan yang sudah melewati tahap preprocessing dimasukkan ke kolom “REVIEW BERSIH”. Berdasarkan Gambar 4.1 ulasan yang diperoleh pada bulan April adalah yang paling sedikit yaitu tiga ulasan, sedangkan pada bulan Mei dan Juni diperoleh berturut-turut sebanyak 1.537 dan 1.501 ulasan. Dan pada bulan Juli diperoleh ulasan paling banyak yaitu 2.537 ulasan.



**Gambar 4. 1** Jumlah Ulasan Tiap Bulan.

Sedangkan jumlah sentimen pada tiap aspek dapat dilihat pada Gambar 4.2. Aspek yang memiliki sentiment negatif paling banyak adalah aspek pelayanan, yaitu sebanyak 1.679 ulasan.



**Gambar 4. 2** Jumlah Sentimen pada Tiap Aspek.

Selanjutnya setiap kelas dari label pada data ulasan divisualisasikan dengan *wordcloud* menggunakan aplikasi voyant tools. Tujuannya adalah untuk melihat kata apa saja yang banyak muncul pada setiap kelas.



**Gambar 4. 3** Wordcloud Label Sentimen (a) Positif dan (b) Negatif.

Berdasarkan Gambar 4.3, *wordcloud* dari ulasan yang memiliki label sentimen positif menunjukkan kata-kata yang banyak muncul diantaranya adalah aplikasi, mudah, bagus dan lainnya. Ulasan yang diberikan memuat kata-kata dari kepuasan pengguna, contohnya adalah “aplikasi jual barang lengkap”. Sedangkan *wordcloud* dari ulasan yang memiliki label sentiment negatif menunjukkan kata-kata yang banyak muncul diantaranya adalah aplikasi, kecewa, memperbaharui dan lainnya. Pada sentiment negatif ini, ulasan yang diberikan pelanggan memuat ungkapan kekecewaan atau ketidakpuasan, contohnya adalah “semenjak memperbaharui aplikasi lambat”.



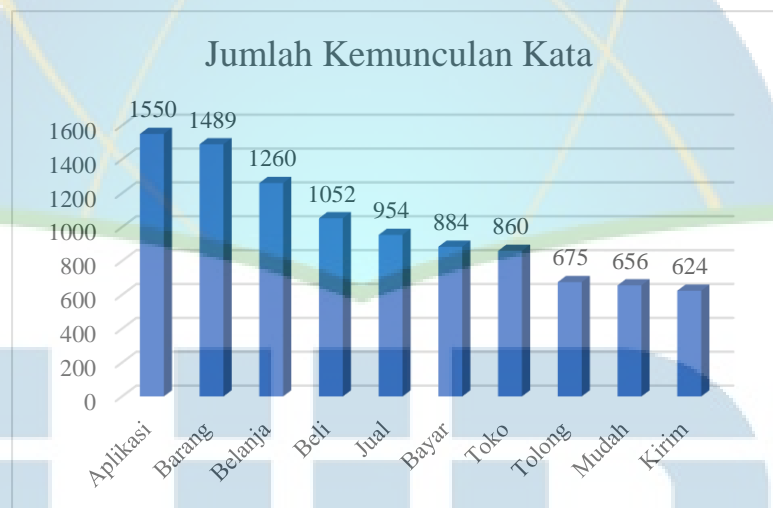
**Gambar 4. 4** Wordcloud Aspek (a) Layanan (b) Sistem (c) Kebermanfaatan.

Sedangkan *wordcloud* untuk ulasan dari tiap aspek dapat dilihat pada Gambar 4.4. Untuk ulasan beraspek layanan, kata yang banyak muncul adalah beli, kirim dan lainnya. Contoh ulasan dari aspek ini adalah “kirim barang lama”. Kemudian untuk ulasan beraspek sistem, kata yang banyak muncul adalah aplikasi, memperbaharui dan

lainnya. Salah satu ulasan dari aspek ini adalah “setelah memperbaharui aplikasi tidak bisa dibuka”. Pada ulasan beraspek kebermanfaatan, kata yang banyak muncul adalah aplikasi, bantu dan lainnya. Salah satu ulasan terkait aspek ini adalah “aplikasinya sangat membantu”.

#### 4.2 Pembobotan Kata

Pembobotan kata menggunakan tfidf. Pada penelitian ini, nilai parameter pada fungsi `Tfidf_vectorizer` yang digunakan adalah sesuai dengan standar modul. Bentuk tfidf yang didapat adalah (5614, 8409), yang artinya bahwa terdapat 5614 baris dokumen dan 8409 kata unik.



**Gambar 4. 5** Jumlah Kemunculan Kata.

Pada Gambar 4.5 menampilkan sepuluh kata dengan jumlah kemunculan terbanyak pada data ulasan yang telah dilakukan *preprocessing*. Kata yang paling banyak muncul adalah kata aplikasi dengan jumlah kemunculan sebanyak 1550. Hal ini menunjukkan bahwa kata tersebut tidak muncul di setiap ulasan sehingga tidak ada kata yang perlu untuk dihilangkan.

### 4.3 Hyperparameter Tunning

Pada pembentukan model sentimen dan aspek dilakukan *cross validation* menggunakan *Grid Search CV*. Tiga kernel yang yaitu linear, poly dan rbf. Parameter yang akan diujikan adalah  $c$  dengan nilai sebagai berikut, yaitu 1, 10, 100 dan 1000. Selain itu, parameter gamma juga diujikan dengan nilai sebagai berikut, yaitu 0.01, 0.1, 1, 10, 100.

**Tabel 4. 2** Hasil *Grid Search CV* Tiap Kernel Sentimen.

Kernel	Parameter Terbaik	Akurasi
Linear	$c=1$	0.696
Poly	$c=1, d=0$	0.692
RBF	$c=1, \text{gamma}=1$	0.676

Tabel 4.2 merupakan hasil *grid search* dari tiap kernel dengan parameter terbaiknya. Untuk model sentimen, nilai akurasi tertinggi didapatkan ketika kernel yang digunakan adalah linear dengan parameternya yaitu  $c=1$ .

**Tabel 4. 3** Hasil *Grid Search CV* Tiap Kernel Aspek.

Kernel	Parameter Terbaik	Akurasi
Linear	$c=1$	0.742
Poly	$c=1000, d=1$	0.682
RBF	$c=1, \text{gamma}=1$	0.731

Tabel 4.3 merupakan hasil *grid search* dari tiap kernel dengan parameter terbaiknya. Untuk model aspek, nilai akurasi tertinggi didapatkan ketika kernel yang digunakan adalah linear dengan parameternya yaitu  $c=1$ .

### 4.4 Hasil Klasifikasi Sentimen dan Aspek

Pada penelitian ini, model yang digunakan untuk klasifikasi sentimen dan aspek adalah SVM. Berdasarkan hasil *hyperparameter tuning* menggunakan *grid search cv*, kernel yang digunakan adalah linear dan nilai  $c=1$ . Hasil confusion matrix untuk kedua klasifikasi tersebut diberikan sebagai berikut:

**Tabel 4. 4** Confusion Matrix Sentimen.

		Nilai Prediksi	
		Negatif	Positif
Nilai aktual	Negatif	671	110
	Positif	231	111

Berdasarkan Tabel 4.4, pada tabel confusion mamenunjukkan bahwa banyak data yang benar terklasifikasi sebagai ulasan negatif adalah sebanyak 671 data. Sedangkan banyak data yang benar terklasifikasi sebagi ulasan positif sebanyak 111 data. Berdasarkan hasil tabel *confusion matrix* diperoleh nilai akurasi dengan menggunakan rumus (2.1) sebagai berikut:

$$\text{Akurasi} = \frac{TN+TP}{TN+FP+TP+FN} = \frac{671+111}{671+110+231+111} = 0.696.$$

Artinya sebesar 69.6% data dapat model klasifikasikan dengan benar pada bagian sentimen.

**Tabel 4. 5** Confusion Matrix Aspek.

		Nilai Prediksi		
		Layanan	Sistem	Kebermanfaatan
Nilai Aktual	Layanan	426	52	28
	Sistem	75	281	32
	Kebermanfaatan	76	26	127

Berdasarkan Tabel 4.5, menunjukkan bahwa banyak data yang benar terklasifikasi layanan sebanyak 426 data, banyak data yang benar terklasifikasi sistem sebanyak 281



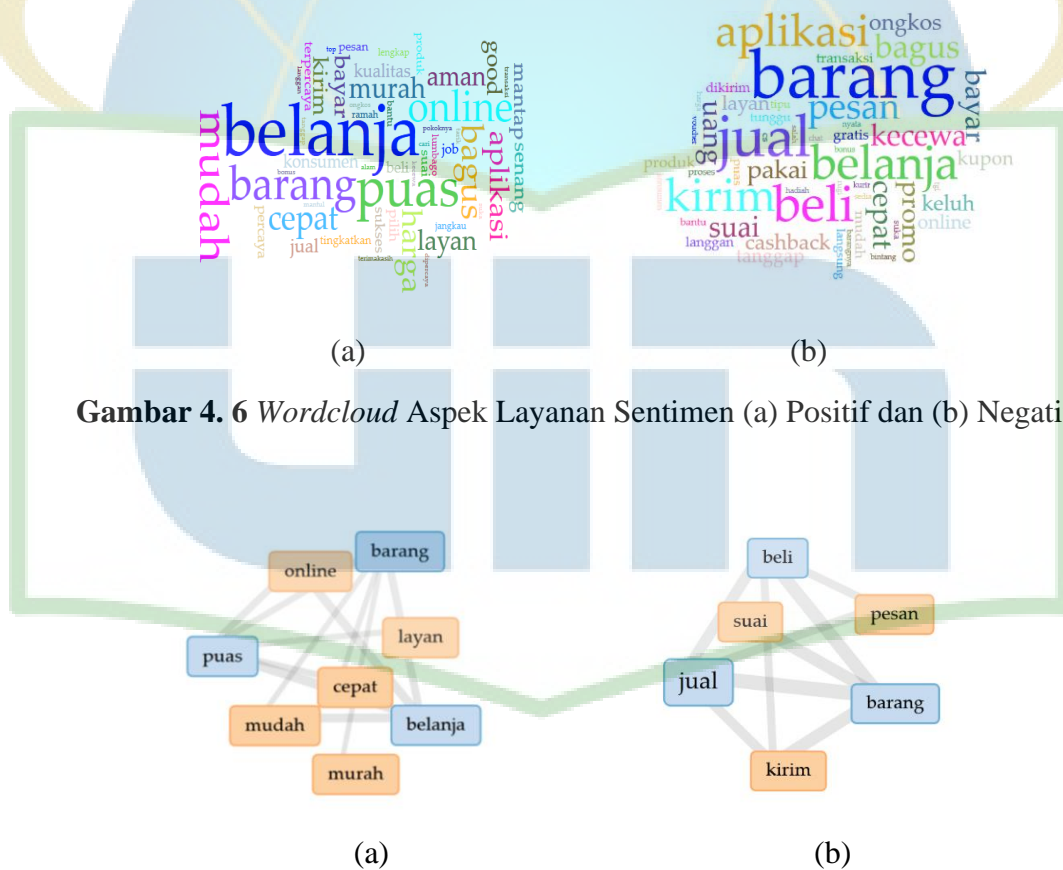
data dan banyak data yang benar terklasifikasi kebermanfaatan sebanyak 127 data. Berdasarkan hasil tabel *confusion matrix* diperoleh nilai akurasi dengan menggunakan rumus (2.1) sebagai berikut:

$$\text{Akurasi} = \frac{TN+TP}{TN+FP+TP+FN} = \frac{426+281+127}{426+281+127+52+28+75+32+76+26} = 0.742.$$

Artinya sebesar 74.2% data dapat model klasifikasikan dengan benar pada bagian aspek.

#### 4.5 Visualisasi dan Interpretasi Data

Visualisasi data menggunakan *wordcloud* dan *wordlink* pada sentimen dari setiap aspek ditunjukkan pada gambar berikut:



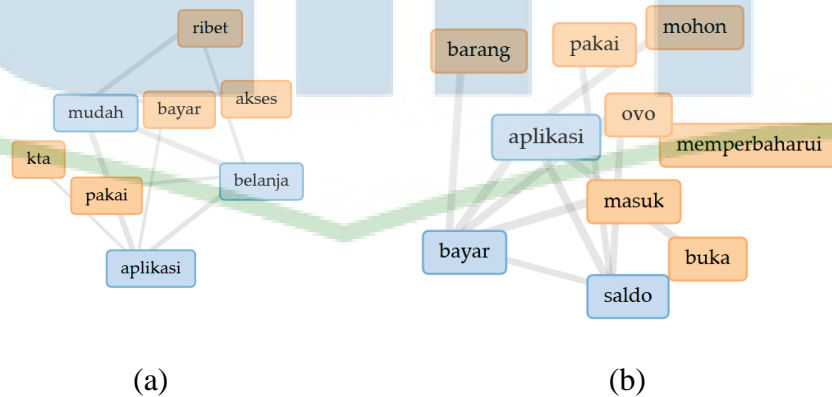
**Gambar 4. 6** Wordcloud Aspek Layanan Sentimen (a) Positif dan (b) Negatif.

**Gambar 4. 7** Wordlink Aspek Layanan Sentimen (a) Positif dan (b) Negatif.

Berdasarkan Gambar 4.6 dan Gambar 4.7, sentimen positif layanan menunjukkan bahwa Tokopedia merupakan aplikasi yang bagus karena para pengguna merasa puas dan aman ketika berbelanja. Barang-barang yang dijualpun berkualitas dan relatif murah. Sedangkan sentimen negatif layanan adalah pelanggan merasa kecewa karena sedikitnya promo, *cashback* dan gratis ongkos kirim. Salah satu contoh ulasannya adalah “Sebenarnya belanja di tokped lebih nyaman, karena klaim garansinya ga rempong kek shopee. Hanya saja tokped jarang ngasih program gratis ongkir sih”.

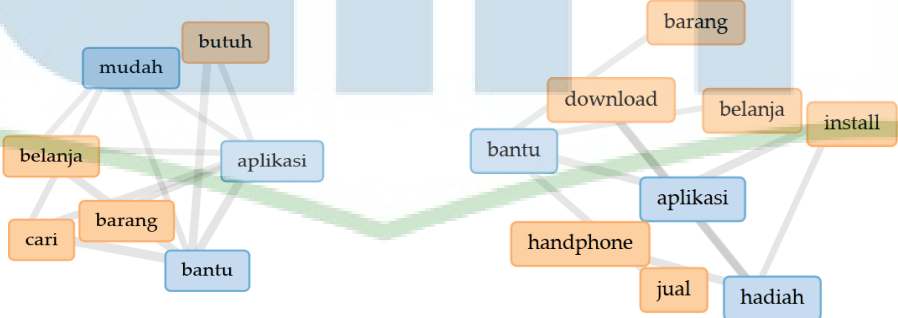
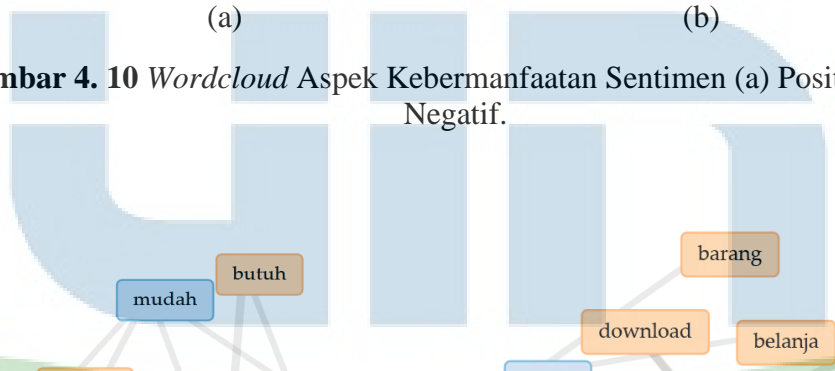


**Gambar 4. 8** Wordcloud Aspek Sistem Sentimen (a) Positif dan (b) Negatif.



**Gambar 4. 9** Wordlink Aspek Sistem Sentimen (a) Positif dan Negatif.

setelah memperbarui aplikasi banyak pelanggan yang gagal membuka aplikasi sehingga beberapa fitur mengalami gangguan seperti foto barang tidak muncul. Salah satu contoh ulasannya adalah “Tolong aplikasi Tokopedia setelah update terbaru karena gambarnya malah tidak muncul hampir semua di produknya, foto review, bahkan foto profil saya juga tidak muncul”.



**Gambar 4. 11** *Wordlink* Aspek Kebermanfaatan Sentimen (a) Positif dan (b) Negatif.

Berdasarkan Gambar 4.10 dan Gambar 4.11, menunjukkan bahwa aplikasi Tokopedia sangat membantu para penggunanya, memudahkan mereka dalam mencari barang yang dibutuhkan. Namun ada juga yang merasa terganggu dengan iklan milik Tokopedia yang muncul saat mereka mengakses situs internet serta mengarahkan untuk memasang aplikasi Tokopedia dengan dijanjikan adanya hadiah padahal tidak ada. Salah satu contoh ulasannya adalah “ga guna, org lg nonton muncul iklan harus download apk nya.. pas udh di download ttp muncul.. GANGGU”.

**Tabel 4. 6** Positif dan Negatif Tiap Aspek

Aspek	Positif	Negatif
Layanan	Murah Aman Cepat	<i>Cashback</i> Ongkos kirim Promo
Sistem	Mudah	Memperbaharui Susah Saldo
Kebermanfaatan	Mudah Membantu Bagus	Iklan Mengganggu Bohong

Pada Tabel 4.8 menjelaskan apa yang menjadi kelebihan serta kekurangan dari tiap aspek pada periode penelitian ini. Hal ini diharapkan dapat dijadikan evaluasi bagi Tokopedia, terlebih pada sisi negatifnya agar terciptanya pengalaman berbelanja yang memuaskan serta menyenangkan bagi para peng

## **BAB V**

### **PENUTUP**

#### **5.1 Kesimpulan**

Data yang digunakan pada penelitian ini adalah 5614 ulasan pengguna aplikasi Tokopedia pada bulan April sampai dengan bulan Juli yang diperoleh melalui proses scraping. Dilakukan pelabelan sentimen yang menghasilkan 3816 ulasan bersentimen negatif dan 1798 ulasan bersentimen positif dan pelabelan aspek yang menghasilkan 2493 ulasan beraspek layanan, 1902 ulasan beraspek sistem dan 1219 ulasan beraspek kebermanfaatan. Proses *preprocessing* yang telah dilakukan diantaranya adalah *case folding*, menghapus simbol, angka dan emotikon, *lemmatization* dan tokenisasi.

Berdasarkan hasil penelitian yang telah dijelaskan, dapat disimpulkan bahwa klasifikasi sentimen dengan menggunakan model SVM menghasilkan nilai akurasi 69,6%. Sedangkan klasifikasi aspek dengan menggunakan model SVM menghasilkan nilai akurasi 74,2%.

Berdasarkan hasil penelitian ini, kelebihan Tokopedia pada aspek layanan adalah pengguna merasa puas dan aman ketika berbelanja. Barang-barang yang dijualpun berkualitas dan relatif murah. Sedangkan kekurangannya adalah sedikitnya promo, *cashback* dan gratis ongkos kirim. Pada aspek sistem kelebihanannya adalah pengguna Tokopedia merasa mudah saat mengakses aplikasi untuk berbelanja, tetapi saat setelah memperbaharui aplikasi beberapa pengguna mengeluhkan sulitnya masuk atau *log in* ke aplikasi. Dan untuk aspek kebermanfaatan aplikasi Tokopedia sangat membantu para penggunanya, memudahkan mereka dalam mencari barang yang dibutuhkan. Namun ada juga yang merasa terganggu dengan iklan milik Tokopedia yang muncul saat mereka mengakses situs internet serta mengarahkan untuk memasang aplikasi Tokopedia dengan dijanjikan adanya hadiah padahal tidak ada.

## 5.2 Saran

Beberapa saran untuk peneliti lainnya yang tertarik untuk melanjutkan penelitian terkait ini diantaranya adalah pengambilan data diurutkan berdasarkan ulasan terbaru dan menambahkan aspek dengan nama “lainnya” untuk ulasan yang tidak termasuk ke dalam aspek yang telah ditentukan. Selain itu, untuk jumlah data yang lebih besar disarankan menggunakan *topic modeling* untuk *text analytic*. Untuk mengetahui performa model, tambahkan perhitungan dari presisi dan *recall* untuk evaluasi model serta lakukan optimasi untuk *recall* sentimen negatif.



## DAFTAR PUSTAKA

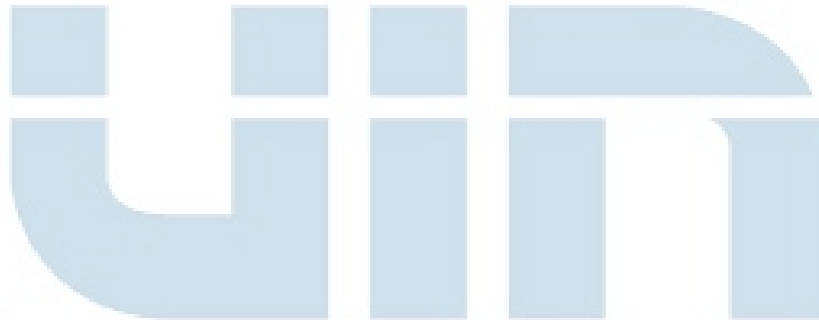
- [1] “Hasil Survei Penetrasi dan Perilaku Pengguna Internet Indonesia 2018,” *APJII*, 2019. [Online]. Available: <https://apjii.or.id/survei>. [Accessed: 24-Jun-2019].
- [2] “Mengungkap Layanan E-Commerce Terpopuler di Indonesia,” *iprice insight*, 2019. [Online]. Available: <https://dailysocial.id/post/mengungkap-layanan-e-commerce-terpopuler-di-indonesia>. [Accessed: 24-Jun-2019].
- [3] S. Gojali and M. L. Khodra, “Aspect based sentiment analysis for review rating prediction,” *4th IGNITE Conf. 2016 Int. Conf. Adv. Informatics Concepts, Theory Appl. ICAICTA 2016*, 2016.
- [4] P. Prameswari, I. Surjandari, and E. Laoh, “Opinion mining from online reviews in Bali tourist area,” *Proceeding - 2017 3rd Int. Conf. Sci. Inf. Technol. Theory Appl. IT Educ. Ind. Soc. Big Data Era, ICSITech 2017*, vol. 2018-Janua, pp. 226–230, 2017, doi: 10.1109/ICSITech.2017.8257115.
- [5] U. Rofiqoh, R. S. Perdana, and M. A. Fauzi, “Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter Dengan Metode Support Vector Machine dan Lexion Based Feature,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 1, no. 12, pp. 1725–1732, 2017.
- [6] T. T. Thet, J. C. Na, and C. S. G. Khoo, “Aspect-based sentiment analysis of movie reviews on discussion boards,” *J. Inf. Sci.*, vol. 36, no. 6, pp. 823–848, 2010, doi: 10.1177/0165551510388123.
- [7] A. Alghunaim, M. Mohtarami, S. Cyphers, and J. Glass, “A Vector Space Approach for Aspect Based Sentiment Analysis,” pp. 116–122, 2015, doi: 10.3115/v1/w15-1516.

- [8] A. Josi, L. A. Abdillah, and Suryayusra, "Penerapan teknik web scraping pada mesin pencari artikel ilmiah," 2014.
- [9] S. N. Kane, A. Mishra, and A. K. Dutta, "Preface: International Conference on Recent Trends in Physics (ICRTP 2016)," *J. Phys. Conf. Ser.*, vol. 755, no. 1, pp. 0–6, 2016, doi: 10.1088/1742-6596/755/1/011001.
- [10] dan H. S. C. Manning, P. Raghavan, "An Introduction and Information Retrieval," *Cambridge Univ. Press*, no. c, 2009.
- [11] J. Wang, S., Lo, D., & Lawall, "Compositional Vector Space Models for Improved Bug Localization," *IEEE*, 2014.
- [12] Howard Anton and Chris Rorres, *Elementary Linear Algebra 11th Edition*. Canada Wiley, 2014.
- [13] M. Awad and R. Khanna, *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*, no. April. 2015.
- [14] I. Syarif, A. Prugel-Bennett, and G. Wills, "SVM parameter optimization using grid search and genetic algorithm to improve classification performance," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 14, no. 4, pp. 1502–1509, 2016, doi: 10.12928/TELKOMNIKA.v14i4.3956.
- [15] A. Indriani, "Klasifikasi Data Forum dengan menggunakan Metode Naive Bayes Classifier," *Semin. Nas. Apl. Teknol. Inf.*, vol. 20, no. ISSN: 1907-5022, pp. 5–10, 2014.
- [16] M. L. Rokach and O, *Data Mining with Decision Tree Theory and Application*, 2nd ed. Singapore: World Scientific, 2015.
- [17] W. A. B. and H. D. Nugroho A S, "Kuliah Umum Ilmu Komputer." [Online]. Available: <http://ilmukomputer.com>. [Accessed: 10-Dec-2019].
- [18] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D.



Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000, doi: 10.1093/bioinformatics/16.10.906.

- [19] F. Rachman and S. W. Purnama, "Perbandingan Klasifikasi Tingkat Keganasan Breast Cancer Dengan Menggunakan Regresi Logistik Ordinal Dan Support Vector Machine ( SVM )," *J. Sains Dan Seni Its*, vol. 1, no. 1, pp. 130–135, 2012.
- [20] S. D. Di and K. Magelang, "Penerapan Metode Klasifikasi Support Vector Machine (Svm) Pada Data Akreditasi Sekolah Dasar (Sd) Di Kabupaten Magelang," *None*, vol. 3, no. 4, pp. 811–820, 2014.
- [21] S. Abtohi, "Implementasi Web Scrapping dan Klasifikasi Sentimen menggunakan Metode Support Vector Machine," UII, 2017.
- [22] E. Prasetyo, "Data Mining Konsep dan Aplikasi Menggunakan matlab," *Yogyakarta:Andi*, 2012.



## LAMPIRAN

### A. Hasil Evaluasi Klasifikasi Sentimen

Label	Precision	Recall	F1-Score
-1	0.74	0.86	0.80
1	0.50	0.32	0.39

### B. Hasil Evaluasi Klasifikasi Aspek

Label	Precision	Recall	F1-Score
Layanan	0,74	0.84	0.79
Sistem	0.78	0.72	0.75
Kebermanfaatan	0,68	0,55	0,61

### C. Klasifikasi Sentimen

#### Memanggil Data yang Digunakan

```
import pandas as pd
dataa=pd.read_csv('REVIEW_PREPROS_EDIT2NEW.csv')
dataa.head()
```

#### Membentuk VSM

```
from sklearn.feature_extraction.text import TfidfVectorizer
#vectorizer = TfidfVectorizer(ngram_range=(1, 2))
vectorizer = TfidfVectorizer()
listdata=dataa['REVIEW BERSIH'].values.astype('object')
listdata = [d for d in listdata]
listdata
vsm_vec = vectorizer.fit(listdata)
```

```
vsm = vsm[vsm.getnnz(1)>0][:,vsm.getnnz(0)>0]
vsm = vsm_vec.transform(listdata)
vsm.shape
```

### **Membagi Data Train dan Test**

```
index = [a for a in range(vsm.shape[0])]
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test, idx_train, idx_test = train_test_split(vsm, y,
index, test_size=0.2 , random_state=0)
print(X_train.shape, X_test.shape)
```

### **Membangun Model Sentimen**

```
import time
import numpy as np
from sklearn.svm import SVR,SVC
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import learning_curve
from sklearn.kernel_ridge import KernelRidge
import matplotlib.pyplot as plt
estimator = SVC(kernel='linear')
param_grid = {"C": [1e0, 1e1, 1e2, 1e3]}
svm = GridSearchCV( estimator=estimator,
                    param_grid=param_grid, n_jobs=-1, refit=True, cv=4,
                    verbose=1)
svm.fit(X_train,y_train)
df=pd.DataFrame(svm.cv_results_)
df
print('best params :{ }\nbest score{ }'.format(svm.best_params_,
svm.best_score_))
```

### **Akurasi**

```
y_pred = svm.predict(X_test)
import sklearn
sklearn.metrics.accuracy_score(y_test,y_pred)
from sklearn.metrics import accuracy_score, confusion_matrix,
classification_report
print('Akurasi = ', accuracy_score(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

### **Menyimpan**

```
def saving_pickle(model,path):
    with open(path,'wb') as file:
        return pickle.dump(model, file)
```

```
#SAVE SPLIT DATA TRAIN TEST
np.save('X_train.npy', X_train)
np.save('y_train.npy', y_train)
np.save('X_test.npy', X_test)
np.save('y_test.npy', y_test)
```

```
#SAVE HASIL VSM
```

```
import pickle
pkl_Filename = 'tfidf_vsm.pkl'
with open(pkl_Filename,'wb') as file:
    pickle.dump(vsm_vec, file)
print(vsm_vec)
```

```
#SAVE MODEL SVM SENTIMEN  
path = "MODEL_SVM_SENTIMEN.pkl"  
saving_pickle(svm,path)
```

```
#SAVE HASIL PREDIKSI SENTIMEN  
prediksi_test.to_csv('prediksi_test.csv')
```

#### D. Klasifikasi Aspek

##### **Memanggil Data yang Digunakan**

```
import pandas as pd  
dataa=pd.read_csv('REVIEW_PREPROS_EDIT2NEW.csv')  
dataa.head()
```

##### **Memanggil VSM yang Telah Disimpan**

```
def load_model_sbr(path):  
    with open(path,'rb') as file:  
        model = pickle.load(file)  
    return model  
  
def saving_pickle(model,path):  
    with open(path,'wb') as file:  
        return pickle.dump(model, file)
```

```
vsm_path = "tfidf_vsm.pkl"  
import pickle  
vsm = load_model_sbr(vsm_path)
```

### **Membagi Data Train dan Test**

```
from sklearn.model_selection import train_test_split
cleanreview = data['REVIEW']
X = vsm.transform(cleanreview)
y = data['ASPECT_LABEL']
X_train, X_test, y_train, y_test, idx_train, idx_test = train_test_split(X, y,
index, test_size=0.2 , random_state=0)
print(X_train.shape, X_test.shape)
```

### **Membangun Model**

```
import time
import numpy as np
from sklearn.svm import SVR,SVC
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import learning_curve
from sklearn.kernel_ridge import KernelRidge
import matplotlib.pyplot as plt

svm = GridSearchCV(SVC(kernel='linear'),
                    param_grid={"C": [1e0, 1e1, 1e2, 1e3]},
                    n_jobs=-1, refit=True, cv=4, verbose=1)
svm.fit(X_train,y_train)
df=pd.DataFrame(svm.cv_results_)
df.to_csv('cv_result_svm_aspect.csv')
df
print('best params : { }\nbest score : { }'.format(svm.best_params_,
svm.best_score_))
```

### **Akurasi**

```
y_pred = svm.predict(X_test)
import sklearn
sklearn.metrics.accuracy_score(y_test,y_pred)
from sklearn.metrics import accuracy_score, confusion_matrix,
classification_report
print('Akurasi = ', accuracy_score(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

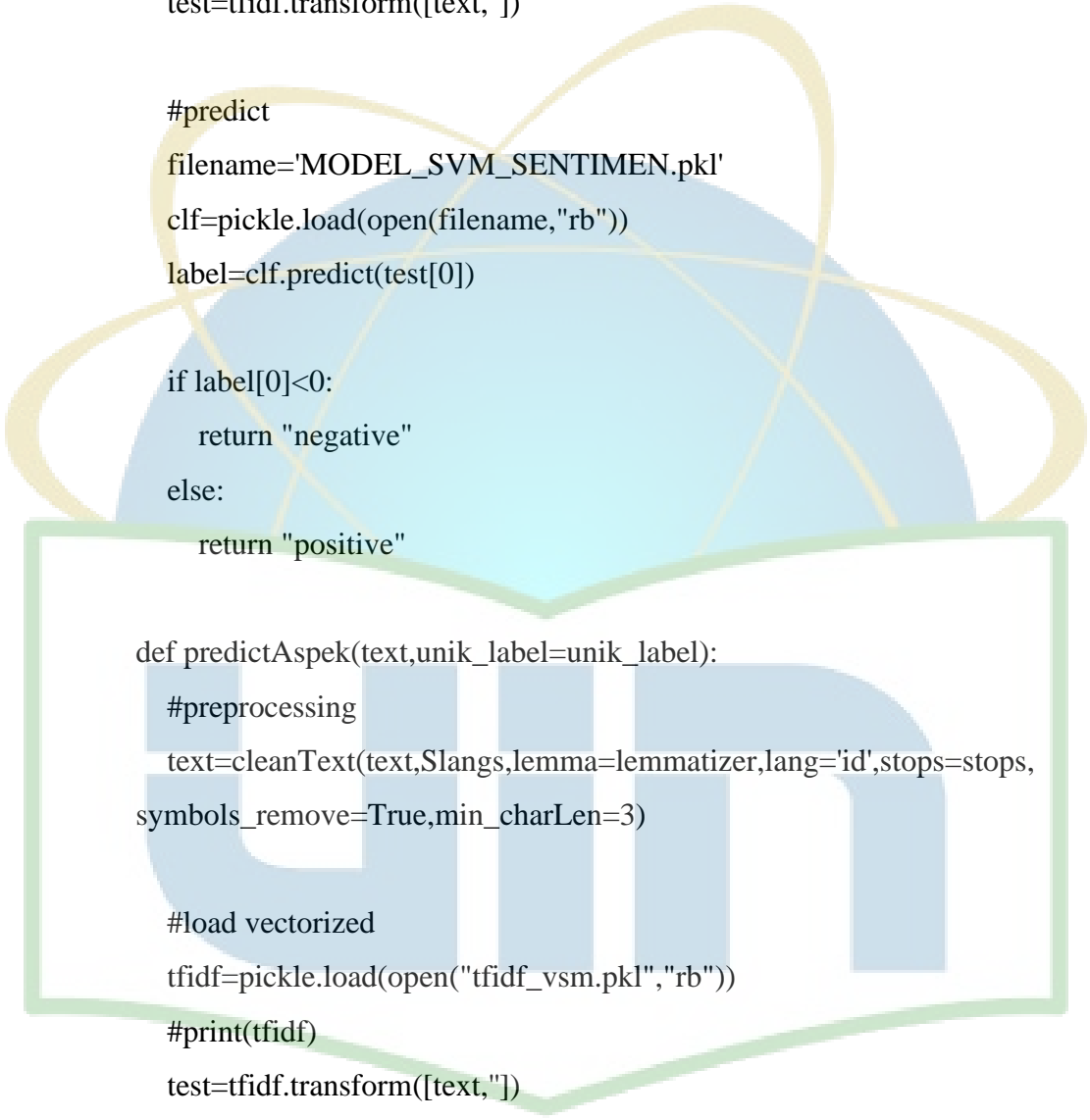
### **Menyimpan**

```
#SAVE VSM ASPEK
np.save('X_train_asp.npy', X_train)
np.save('y_train_asp.npy', y_train)
np.save('X_test_asp.npy', X_test)
np.save('y_test_asp.npy', y_test)

#SAVE MODEL SVM ASPEK
path = 'MODEL_SVM_ASPECT.pkl'
saving_pickle(svm,path)
```

### **Uji Model**

```
Slangs = LoadSlang('C:/WinPython/notebooks/SKRIPSI/slang.txt')
stops, lemmatizer = LoadStopWords(lang='id')
def predictSenti(text):
    #preprocessing
    text=cleanText(text,Slangs,lemma=lemmatizer,lang='id',stops=stops,
symbols_remove=True,min_charLen=3)
```



```

#load vectorized
tfidf=pickle.load(open("tfidf_vsm.pkl","rb"))
#print(tfidf)
test=tfidf.transform([text,"])

#predict
filename='MODEL_SVM_SENTIMEN.pkl'
clf=pickle.load(open(filename,"rb"))
label=clf.predict(test[0])

if label[0]<0:
    return "negative"
else:
    return "positive"

def predictAspek(text,unik_label=unik_label):
    #preprocessing
    text=cleanText(text,Slangs,lemma=lemmatizer,lang='id',stops=stops,
symbols_remove=True,min_charLen=3)

    #load vectorized
    tfidf=pickle.load(open("tfidf_vsm.pkl","rb"))
    #print(tfidf)
    test=tfidf.transform([text,"])

#predict
filename='MODEL_SVM_ASPECT.pkl'
clf=pickle.load(open(filename,"rb"))
label=clf.predict(test[0])

```



```
return unik_label[label][0]
```

```
def predict_text(text, unik_label=unik_label):  
    sentimen = predictSenti(text)  
    aspek = predictAspek(text)  
  
    return {'sentimen':sentimen, 'aspek':aspek}  
print(predict_text('ini aplikasi jelek banget'))
```

