KLASIFIKASI KUALITAS ULASAN PRODUK BERDASARKAN *SEMANTIC* DAN *STRUCTURAL FEATURES* MENGGUNAKAN *SUPPORT VECTOR MACHINE*

SKRIPSI

Disusun Oleh:

ILHAM AKHYAR FIRDAUS
H06218013

PROGRAM STUDI SISTEM INFORMASI
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI SUNAN AMPEL
SURABAYA
2022

PERNYATAAN KEASLIAN

Saya yang bertanda tangan di bawah ini,

Nama	:	ILHAM AKHYAR FIRDAUS
NIM	:	H06218013
Program Studi	:	Sistem Informasi
Angkatan	:	2018

Menyatakan bahwa saya tidak melakukan plagiat dalam penulisan skripsi saya yang berjudul: "KLASIFIKASI KUALITAS ULASAN PRODUK BERDASARKAN SEMANTIC DAN STRUCTURAL FEATURES MENGGUNAKAN SUPPORT VECTOR MACHINE". Apabila suatu saat nanti terbukti saya melakukan tindakan plagiat, maka saya bersedia menerima sanksi yang telah ditetapkan.

Demikian pernyataan keaslian ini saya buat dengan sebenar-benarnya.

Surabaya, 12 Agustus 2022
Yang menyatakan,

Ilham Akhyar Firdaus
NIM. H06218013

LEMBAR PERSETUJUAN PEMBIMBING

Skripsi oleh

NAMA	:	ILHAM AKHYAR FIRDAUS
NIM	:	H06218013
JUDUL	:	KLASIFIKASI KUALITAS ULASAN PRODUK BERDASARKAN *SEMANTIC* DAN S*TRUCTURAL FEATURES* MENGGUNAKAN *SUPPORT VECTOR MACHINE*

Ini telah diperiksa dan disetujui untuk diujikan.

Surabaya, 08 Agustus 2022

Dosen Pembimbing 1	Dosen Pembimbing 2
Dwi Kolliawati, MT	Dr. Eng. Anang Kunaefi, M.Kom
NIP. 197909272014032001	NIP. 197911132014031001

PENGESAHAN TIM PENGUJI SKRIPSI

Skripsi Ilham Akhyar Firdaus ini telah dipertahankan
di depan tim penguji skripsi
di Surabaya, 11 Agustus 2022

Mengesahkan,
Dewan Penguji

Dosen Penguji 1

Mujib Ridwan, S.Kom., M.T
NIP. 198604272014031004

Dosen Penguji 2

Bayu Adhi Nugroho, Ph.D.
NIP. 197905182014031001

Dosen Penguji 3

Dwi Rolliawati, MT
NIP. 197909272014032001

Dosen Penguji 4

Dr. Eng. Anang Kunaefi, M.Kom
NIP. 197911132014031001

Mengetahui,
Dekan Fakultas Sains dan Teknologi
UIN Sunan Ampel Surabaya

Dr. A. Saepul Hamdani, M.Pd
NIP. 196507312000031002

KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN AMPEL SURABAYA
PERPUSTAKAAN
Jl. Jend. A. Yani 117 Surabaya 60237 Telp. 031-8431972 Fax.031-8413300
E-Mail: perpus@uinsby.ac.id

LEMBAR PERNYATAAN PERSETUJUAN PUBLIKASI
KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika UIN Sunan Ampel Surabaya, yang bertanda tangan di bawah ini, saya:

Nama : ILHAM AKHYAR FIRDAUS
NIM : H06218013
Fakultas/Jurusan : SAINS DAN TEKNOLOGI/SISTEM INFORMASI
E-mail address : ilhamakhyar202@gmail.com

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Perpustakaan UIN Sunan Ampel Surabaya, Hak Bebas Royalti Non-Eksklusif atas karya ilmiah :
■ Sekripsi ☐ Tesis ☐ Desertasi ☐ Lain-lain (.....)
yang berjudul :
KLASIFIKASI KUALITAS ULASAN PRODUK BERDASARKAN SEMANTIC DAN STRUCTURAL FEATURES MENGGUNAKAN SUPPORT VECTOR MACHINE

beserta perangkat yang diperlukan (bila ada). Dengan Hak Bebas Royalti Non-Ekslusif ini Perpustakaan UIN Sunan Ampel Surabaya berhak menyimpan, mengalih-media/format-kan, mengelolanya dalam bentuk pangkalan data (database), mendistribusikannya, dan menampilkan/mempublikasikannya di Internet atau media lain secara ***fulltext*** untuk kepentingan akademis tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan atau penerbit yang bersangkutan.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak Perpustakaan UIN Sunan Ampel Surabaya, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini yang saya buat dengan sebenarnya.

Surabaya, 12 Agustus 2022

Penulis

(ILHAM AKHYAR FIRDAUS)

ABSTRAK

KLASIFIKASI KUALITAS ULASAN PRODUK BERDASARKAN *SEMANTIC* DAN *STRUCTURAL FEATURES* MENGGUNAKAN *SUPPORT VECTOR MACHINE*

Oleh:

Ilham Akhyar Firdaus

Ulasan produk merupakan opini tertulis yang disampaikan oleh konsumen dalam menilai suatu produk. Adanya ulasan produk menjadi penting dikarenakan dapat membantu konsumen dalam membuat keputusan pada pembelian produk yang lebih baik. Namun ulasan produk dapat menjadi tidak penting apabila kualitas informasi dari ulasan tersebut tidak bermanfaat. Hal tersebut dapat diminimalisir apabila dilakukan klasifikasi untuk mengetahui ulasan mana yang bermanfaat atau tidak bermanfaat. Agar hal tersebut dapat tercapai, maka pada penelitian ini diterapkan model *support vector machine* (SVM) berbasiskan pada pengenalan pola menggunakan *semantic* dan *structural features* untuk dapat melakukan klasifikasi pada teks ulasan berdasarkan karakteristiknya. Hasil akhir menunjukan bahwa model SVM pada *semantic feature* memiliki nilai *f-measure* tertinggi sebesar 0.825. Sedangkan pada *structural feature* nilai *f-measure* tertinggi adalah sebesar 0.823. Dari hal tersebut dapat disimpulkan bahwa *semantic feature* dapat digunakan untuk mengidentifikasi karakteristik dari teks ulasan yang bermanfaat atau tidak bermanfaat dengan baik.

Kata Kunci: Ulasan Produk, Klasifikasi Teks, *Support Vector Machine*, *Semantic Feature*, *Structural Feature*

ABSTRACT

CLASSIFICATION OF PRODUCT REVIEW QUALITY BASED ON SEMANTIC AND STRUCTURAL FEATURES USING SUPPORT VECTOR MACHINE

***By*:**

Ilham Akhyar Firdaus

Product reviews are written opinions submitted by consumers in assessing a product. The existence of product reviews is important because it can help consumers in making better product purchasing decisions. But product reviews can become unimportant if the quality of the information from those reviews is not helpful. This can be minimized if classification is carried out to find out which reviews are useful or not. For this to be achieved, this research applies a support vector machine (SVM) model based on pattern recognition using semantics and structural features to be able to classify the review text based on its characteristics. The final result shows that the SVM model on the semantic feature has the highest f-measure value of 0.825. Meanwhile, in the structural feature, the highest f-measure value is 0.823. From this, it can be concluded that semantic features can be used to identify the characteristics of the review text that are useful or not useful properly.

Keywords: *Product Reviews, Text Classification, Support Vector Machine, Semantic Feature, Structural Feature*

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi telah menghasilkan banyak perubahan serta kemajuan sosial dalam beberapa puluh tahun terakhir, sehingga mengubah perilaku dan hubungan sosial pada masyarakat (Moreno-Llamas, García-Mayor, & De la Cruz-Sánchez, 2020). Salah satu perubahan perilaku masyarakat saat ini adalah adanya perubahan pada perilaku gaya hidup. Perubahan perilaku gaya hidup tersebut muncul dikarenakan adanya intensitas informasi yang bermacam-macam pada kehidupan sehari-hari. Salah satu contohnya adalah akibat dari penggunaan aplikasi *mobile*. Kemudahan dalam menggunakan aplikasi *mobile* telah dirasakan oleh masyarakat di berbagai bidang. Seperti pada bidang bisnis, kemudahan yang dapat dirasakan oleh masyarakat saat ini adalah dapat melakukan transaksi jual beli secara *online* melalui *platform* seperti *e-commerce*.

Transaksi jual beli dengan menggunakan *e-commerce* telah menjadi perilaku gaya hidup yang umum di masyarakat dikarenakan masyarakat dapat membeli produk tanpa mengunjungi toko secara fisik yang dapat mengurangi upaya untuk berbelanja secara *offline* sehingga dapat menghemat waktu serta memungkinkan untuk mengenali produk dengan lebih cepat. Sehingga fokus dari *e-commerce* adalah untuk memaksimalkan efisiensi belanja dengan strategi seperti kemudahan dalam pencarian produk, pembelian dengan sekali klik, dan katalog *virtual* yang memiliki spesifikasi dan rekomendasi berdasarkan perilaku belanja konsumen di masa lalu (Z. Huang & Benyoucef, 2013). Hal tersebut mengakibatkan konsumen dari *e-commerce* mengalami peningkatan serta terjadinya *electronic word of mouth*.

Electronic word of mouth (eWOM) merujuk pada umpan balik dan sudut pandang konsumen terkait produk atau servis yang dapat berupa *vote like*, komentar, rating, ulasan, video testimoni, gambar, atau sebuah postingan pada blog (Donthu et al., 2021). Dikarenakan adanya eWOM, konsumen lebih tertarik pada produk yang dibahas secara *online* dibandingkan dengan produk yang dibahas

secara tradisional yang mana eWOM lebih kaya akan sumber informasi yang objektif (Pai et al., 2013). Salah satu jenis dari eWOM adalah ulasan produk.

Ulasan produk menjadi penting dikarenakan berefek pada keputusan konsumen dalam pembelian produk berdasarkan atribut, situasi penggunaan, dan performa produk berdasarkan konsumen yang lain (Weisstein et al., 2017). Meskipun begitu, ulasan produk juga dapat menjadi tidak penting apabila kualitas informasi dari ulasan tersebut tidak bermanfaat. Untuk dapat mengetahui apakah ulasan tersebut bermanfaat atau tidak adalah dengan cara mengenali pola dari teks ulasan yang bermanfaat menggunakan *features* yang ada seperti panjang atau tingkat keterbacaan ulasan (Chua & Banerjee, 2016a). Adapun konsumen dapat mengetahui dengan mudah yaitu apakah ulasan tersebut bermanfaat atau tidak adalah dengan cara melihat jumlah *vote* manfaat yang ada pada tiap ulasan (Cao, Duan, & Gan, 2011).

Vote manfaat yang ada pada tiap ulasan dapat menentukan apakah ulasan tersebut bermanfaat atau tidak dikarenakan ulasan tersebut telah dinilai kemanfaatannya oleh konsumen yang lain dengan cara melakukan *voting* manfaat pada ulasan tersebut. Dengan banyaknya variasi pada jumlah dan informasi dari suatu ulasan pada tiap produk, maka penting untuk mengetahui apakah ulasan tersebut bermanfaat atau tidak secara cepat sehingga calon konsumen dapat dengan mudah menilai apakah produk tersebut sudah sesuai dengan apa yang dibutuhkan atau tidak berdasarkan pada ulasan-ulasannya. Hal itu dapat diterapkan dengan cara melakukan klasifikasi pada ulasan bermanfaat atau tidak dengan berdasarkan pada *feature* ulasan serta melakukan tahapan *preprocessing* yang tepat.

Dari berbagai macam model klasifikasi yang ada, model klasifikasi menggunakan *support vector machine* (SVM) merupakan model yang dapat melakukan klasifikasi pada data berupa teks (Wei, Wei, & Wang, 2012). Penelitian Khorsheed & Al-Thubaity (2013) yang menggunakan data teks berupa *dataset arabic* mencoba untuk membandingkan tujuh model klasifikasi (C4.5, C5.0, *multi layer perceptron neural network*, SVM, *naive bayes*, *k-nearest neighbors*) menyimpulkan bahwa model klasifikasi SVM memiliki performa klasifikasi yang lebih baik dibandingkan dengan ke-enam model lainnya diikuti oleh C4.5 dan *naive bayes*. Model klasifikasi SVM mengimplementasikan prinsip *structural risk*

minimization (SRM) ketika melakukan klasifikasi yang digunakan untuk menemukan *hyperplane* pemisah yang optimal, hal tersebut mengakibatkan nilai akurasi dari hasil klasifikasi pada SVM memiliki nilai yang cukup tinggi (Wan et al., 2012). Tidak hanya itu model SVM juga mampu menangani input data yang berdimensi besar seperti halnya data teks (Balakumar & Mohan, 2019). Sehingga pada penelitian ini akan digunakan model SVM untuk melakukan klasifikasi pada data teks berdasarkan *features* yang tersedia.

Features pada teks merupakan sekumpulan individual item yang berfungsi untuk mengetahui bagaimana karakteristik dari suatu teks. Sehingga ulasan yang bermanfaat dapat diidentifikasi melalui *features* yang tersedia pada ulasan tersebut. Beberapa *features* seperti *structural feature, semantic feature, syntactic feature, lexical feature* dan *review metadata* pada ulasan dapat digunakan untuk mengevaluasi apakah ulasan tersebut bermanfaat atau tidak (Dash, Zhang, & Zhou, 2021).

Semantic feature merupakan *feature* yang berperan untuk mengetahui makna kata pada suatu teks (Kim et al., 2018). *Semantic feature* merupakan *feature* yang paling penting dan berpengaruh dalam mendapatkan ulasan yang bermanfaat dibandingkan dengan beberapa *features* yang lain (Ngo-Ye & Sinha, 2014). Terlebih *semantic feature* menghasilkan kinerja klasifikasi yang lebih akurat jika dibandingkan dengan teks klasifikasi secara tradisional seperti representasi vektor sederhana dari kata atau representasi data menggunakan *bag-of-words* (Altınel & Ganiz, 2018). Terdapat beberapa teknik dalam mengekstrak *semantic feature* seperti *n-gram, latent dirichlet allocation, skip gram negative sampling, term frequency-inverse document frequency* (TF-IDF) dan *global vector*. Dari beberapa teknik tersebut, penggunaan TF-IDF memiliki nilai akurasi klasifikasi yang tinggi jika dibandingkan dengan beberapa metode lainnya (Du et al., 2019).

Structural feature merupakan *feature* yang berperan untuk mengetahui struktur dan format pada sebuah dokumen teks seperti jumlah kata pada tiap kalimat, panjang ulasan, dan jumlah karakter. Perhitungan jumlah kata pada tiap kalimat dapat menjadi variabel penduga yang kuat untuk menghasilkan nilai akurasi klasifikasi yang tinggi ketika mencapai jumlah kata tertentu (A. H. Huang et al.,

2015). Kedalaman dari ulasan yaitu dari banyaknya jumlah kata yang ada juga memiliki korelasi positif terhadap penilaian ulasan yang bermanfaat (Wu, 2017).

Pada beberapa penelitian sebelumnya (Pan & Zhang, 2011) (Lee & Choeh, 2014) masih berfokus pada penggunaan satu *feature* untuk mengidentifikasi karakteristik teks ulasannya. Berdasarkan paparan penelitian dari (Meng et al., 2021) pada penggunaan *structural feature* terdapat rekomendasi untuk tidak hanya melihat dari sisi fitur karakter atau kata dari sebuah teks ulasan saja, tetapi juga mempertimbangkan hubungan atau makna dari karakter atau kata itu sendiri. Dari hal tersebut, maka pada penelitian ini akan berfokus pada penggunaan serta penggabungan dua *features* untuk mengetahui bagaimana pengaruhnya dalam memprediksi ulasan mana yang bermanfaat atau tidak bermanfaat sehingga dapat menghasilkan nilai akurasi yang tinggi. Dari uraian latar belakang tersebut, maka penelitian yang akan dilakukan berjudul "KLASIFIKASI KUALITAS ULASAN PRODUK BERDASARKAN *SEMANTIC* DAN *STRUCTURAL FEATURES* MENGGUNAKAN *SUPPORT VECTOR MACHINE*". Diharapkan dengan adanya penelitian ini dapat mengetahui *features* mana dan teknik *preprocessing* seperti apa yang dapat dilakukan untuk menghasilkan performa model klasifikasi yang akurat dalam mengetahui ulasan mana yang bermanfaat atau tidak bermanfaat.

1.2 Perumusan Masalah

Berdasarkan uraian dari latar belakang tersebut, maka pada penelitian ini dapat dibuat perumusan masalah sebagai berikut:

1. Bagaimana melakukan ekstraksi *semantic feature* untuk klasifikasi ulasan bermanfaat?
2. Bagaimana melakukan ekstraksi *structural feature* untuk klasifikasi ulasan bermanfaat?
3. Bagaimana performa model *support vector machine* dalam melakukan klasifikasi ulasan bermanfaat?

1.3 Batasan Masalah

Adanya batasan masalah dibuat untuk menghindari penelitian yang melebar sehingga lebih fokus pada topik yang akan diteliti. Batasan masalah dari penelitian ini adalah sebagai berikut:

1. Data yang digunakan dalam penelitian ini adalah *dataset* terbuka dari ulasan produk amazon berbahasa inggris tahun 2015.
2. Nilai *threshold* yang digunakan untuk mengidentifikasi ulasan bermanfaat adalah 0.6.

1.4 Tujuan Penelitian

Berdasarkan uraian latar belakang dan perumusan masalah, maka ditentukan tujuan dari penelitian ini adalah sebagai berikut:

1. Dapat mengetahui *features* mana yang dapat menjadi variabel penduga terbaik untuk ulasan bermanfaat.
2. Dapat mengetahui hasil performa model *support vector machine* terhadap klasifikasi ulasan bermanfaat.
3. Dapat mengetahui pengaruh tahapan *preprocessing* yang tepat pada proses klasifikasi ulasan bermanfaat.

1.5 Manfaat Penelitian

Penelitian yang dilakukan diharapkan dapat memberikan manfaat secara langsung atau tidak langsung. Manfaat yang diharapkan pada penelitian ini adalah sebagai berikut:

1. Secara Akademis
    a. Dapat menjadi sumbangan ilmiah dalam penelitian mengenai klasifikasi ulasan produk terlebih pada ulasan bermanfaat dengan berdasarkan pada *semantic* dan *structural feature*.
    b. Sebagai referensi serta wawasan tambahan dalam melakukan penelitian mengenai klasifikasi ulasan produk mulai tahap awal hingga tahap akhir.

2. Secara Aplikatif

Memberikan manfaat berdasarkan hasil serta pembahasan pada penelitian yaitu terkait bagaimana tahapan dari klasifikasi teks ulasan, terlebih untuk mengetahui apakah ulasan tersebut bermanfaat atau tidak dengan cara mengenali pola dari teks ulasan tersebut.

BAB II
TINJAUAN PUSTAKA

2.1 Tinjauan Penelitian Terdahulu

Tinjauan penelitian terdahulu adalah mempelajari penelitian yang sudah dilakukan sebelumnya yang relevan pada penelitian ini untuk digunakan sebagai memperdalam pemahaman terkait hasil dari penelitian tersebut. Dipaparkan juga korelasi penelitian ini dengan penelitian terdahulu sebagai bahan referensi untuk mengetahui persamaan dan perbedaannya. Berikut pada Tabel 2.1 merupakan tinjauan dari beberapa penelitian terdahulu.

Tabel 2.1 Tinjauan Penelitian Terdahulu

No.	Judul	Hasil	Korelasi
1.	*Comparison of Supervised Classification Models on Textual Data* *(Hsu, 2020)	Model klasifikasi SVM memiliki nilai akurasi yang tinggi dibandingkan dengan tujuh model klasifikasi lainnya dikarenakan SVM mampu memetakan teks secara linear.	Model SVM dapat diterapkan pada data teks dengan nilai akurasi klasifikasi yang tinggi.
2.	*Sentimental text mining based on an additional features method for text classification* *(Cheng & Chen, 2019)	Modifikasi kata pada *dataset* ulasan film dengan melakukan berbagai teknik *preprocessing* pada model SVM, jika dibandingkan dengan tanpa melakukan modifikasi dapat menghasilkan *accuracy* model yang lebih baik.	Nilai akurasi dari model SVM dapat dipengaruhi oleh tahapan *preprocessing* pada kata.
3.	*Feature selection using an improved Chi-square for Arabic text classification* *(Bahassine et al., 2020)	Pemilihan *feature selection* dan jumlah *feature* yang sesuai pada data teks menghasilkan peningkatan nilai akurasi model SVM dan *decision tree*.	Nilai akurasi model klasifikasi dapat dipengaruhi oleh pemilihan *feature selection* dan jumlah *feature* yang tepat terutama pada model. SVM
4.	*A feature-centric spam email detection model using diverse supervised machine learning*	Hasil analisis pada lima *feature* yang digunakan yaitu *content, sentiment, semantic, user*, dan *lexicon*	*Semantic feature* dapat digunakan sebagai variabel penduga yang baik pada model

	algorithms *(Zamir et al., 2020)	pada klasifikasi *spam* email menghasilkan bahwa *sentiment* dan *semantic feature* memiliki nilai penduga yang paling kuat dibandingkan dengan *feature* lainnya.	klasifikasi.
5.	*Does the review deserve more helpfulness when its title resembles the content? Locating helpful reviews by text mining* *(Zhou et al., 2020)	Hasil klasifikasi model *logistic regression* tidak jauh beda dengan model *tobit regression* ketika diberi nilai *threshold* untuk ulasan bermanfaat sebesar 0.6	Nilai *threshold* 0.6 dapat digunakan untuk membedakan ulasan yang bermanfaat yaitu apabila lebih dari 0.6 maka ulasan tersebut bermanfaat dan juga sebaliknya
6.	*An analysis of review content and reviewer variables that contribute to review helpfulness* *(M. S.I. Malik & Hussain, 2018)	Hasil dari membandingkan beberapa variabel untuk klasifikasi ulasan bermanfaat, variabel isi ulasan (*review content*) merupakan indikator yang paling efektif dibandingkan dengan lainnya.	*Feature* isi ulasan seperti *semantic* dan *structural* dapat digunakan sebagai indikator ulasan bermanfaat.
7.	*Understanding the effects of different review features on purchase probability* *(S. J. Kim, Maslowska, & Malthouse, 2018)	Panjang ulasan memiliki korelasi positif pada keputusan pembelian apabila mencapai banyak kata sekitar 20-55 dan menjadi negatif apabila ulasan terlalu panjang (lebih dari 55 kata).	*Structural features* seperti panjang kata pada ulasan dapat dijadikan sebagai indikator dalam melakukan klasifikasi pada ulasan yang bermanfaat.
8.	*The Effects of Credible Online Reviews on Brand Equity Dimensions and Its Consequence on Consumer Behavior* *(Chakraborty & Bhat, 2018)	Kualitas dan konsistensi ulasan mempengaruhi tingkat kepercayaan pada ulasan yang mengakibatkan calon pembeli dapat percaya pada produk/brand tersebut.	Dengan melakukan klasifikasi pada ulasan yang bermanfaat dapat mempengaruhi keputusan pembeli dalam membeli produk dikarenakan pembeli dapat menilai kualitas dari produk tersebut dengan berdasarkan pada ulasan yang ada.

Dari beberapa penelitian yang dirangkumkan pada Tabel 2.1 tersebut dapat disimpulkan bahwa model SVM dapat digunakan untuk melakukan klasifikasi pada data teks. Nilai akurasi pada model SVM juga bernilai tinggi dengan dapat

dipengaruhi oleh beberapa hal yaitu seperti tahapan *preprocessing*, pemilihan *feature selection*, serta jumlah *feature* yang tepat. Adapun untuk dapat mengidentifikasi ulasan produk yang bermanfaat dapat dilakukan ekstraksi *semantic* dan *structural feature* pada teks dengan melihat threshold ulasan bermanfaat yaitu apabila lebih dari 0.6 maka ulasan tersebut merupakan ulasan bermanfaat begitu juga dengan sebaliknya.

2.2 Teori-teori Dasar

Terkait dengan konsep, teori, dan variabel yang digunakan pada penelitian ini. Maka pada sub ini akan dijelaskan semua hal tersebut seperti *e-commerce*, ulasan produk, *helpful votes, automatic spelling correction*, *text mining*, *text preprocessing*, *support vector machine*, *feature extraction*, *confusion matrix*, dan *k-fold cross validation*.

2.2.1 E-commerce

Istilah dari *electronic commerce* (*e-commerce*) mengacu pada sebuah model bisnis yang memungkinkan perusahaan, kelompok, serta individu untuk dapat membeli atau menjual barang atau jasa melalui internet. Faktor kemudahan merupakan salah satu hal penting yang ditawarkan dari sebuah *e-commerce*. Seperti calon konsumen yang mendapatkan manfaat dari adanya diskon, pencarian yang cepat, banyaknya pilihan jenis produk, opsi pembayaran yang lengkap, dll. (Salehi et al., 2012). Akibat dari hal tersebut *e-commerce* berkembang secara pesat dalam beberapa tahun terakhir.

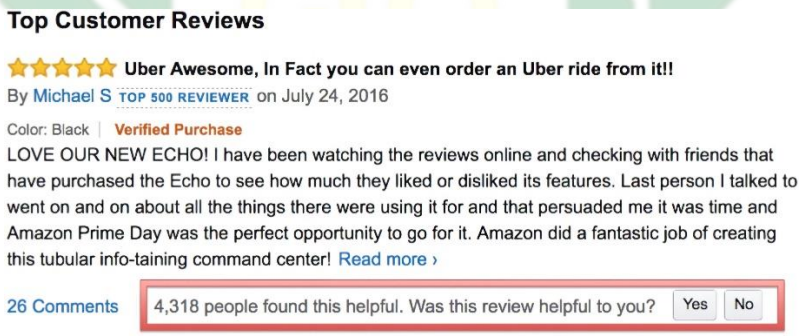
2.2.2 Ulasan Produk

Ulasan produk merupakan salah satu bagian dari *electronic word of mouth* (eWOM) yang merujuk pada ulasan tekstual dari konsumen yang mendeskripsikan terkait karakteristik seperti kelebihan atau kelemahan dari produk tersebut (Lackermair, Kailer, & Kanmaz, 2013). Tujuan utama dari adanya ulasan produk adalah untuk membantu konsumen dalam membuat keputusan pembelian yang lebih baik (Chou, Picazo-Vela, & Pearson, 2013). Dikarenakan ulasan produk merupakan salah satu bagian dari eWOM maka karakteristik yang dimiliki oleh ulasan produk adalah pada pernyataannya yang tidak memiliki niat untuk

melakukan promosi pada produk tersebut, hal ini menjadikan calon konsumen lain lebih percaya akan kredibilitas yang dimiliki oleh ulasan dibandingkan dengan iklan komersial dalam menilai suatu produk (Herr, Kardes, & Kim, 1991).

2.2.3 Helpful Votes

Suatu ulasan dapat dikatakan sebagai ulasan yang bermanfaat apabila ulasan tersebut dapat memfasilitasi proses keputusan dalam pembelian produk pada konsumen (Mudambi & Schuff, 2010). Salah satu indikator dari ulasan yang bermanfaat adalah dengan cara melihat jumlah *vote* manfaat (*helpful votes*) yang ada pada ulasan tersebut. Pengertian dari *helpful votes* merupakan *vote* yang diberikan oleh konsumen lain kepada konsumen yang memberikan ulasan tersebut dikarenakan percaya bahwa ulasan yang diberikan telah membantunya dalam membuat keputusan pembelian terkait produk yang akan dibeli. Berikut pada Gambar 2.1 merupakan contoh dari *helpful votes* pada amazon.



Gambar 2.1 *Helpful Votes* Amazon

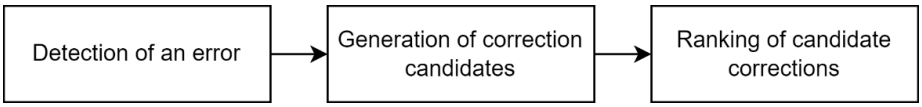
(Alzu'Bi et al., 2019)

Semakin banyak jumlah *helpful votes* yang ada pada ulasan tersebut maka semakin besar juga kemungkinan bahwa ulasan tersebut bermanfaat (Schuckert, Liu, & Law, 2016).

2.2.4 Automatic Spelling Correction

Automatic spelling correction (ASC) merupakan proses dalam melakukan koreksi pada ejaan kata yang salah secara otomatis, dimana ejaan kata yang salah tersebut dapat membuat dokumen teks menjadi sulit dibaca serta sulit untuk diproses sehingga perlu dilakukan koreksi pada ejaannya. Ejaan kata yang salah

juga dapat mengakibatkan menurunnya nilai dari informasi dalam suatu dokumen teks. Secara umum terdapat tiga tahapan dalam ASC (Hládek, Staš, & Pleva, 2020) yaitu sesuai pada Gambar 2.2 berikut.



Gambar 2.2 Tahapan ASC

Gambar 2.2 pada *error detection* merupakan tahapan dalam mendeteksi adanya ejaan yang salah atau tidak pada suatu kata. Terdapat dua kategori *error* yaitu pertama kata tersebut dieja salah namun secara struktur katanya ada di dalam kamus kata yang benar dan yang kedua adalah kata tersebut dieja salah namun tidak ada di kamus kata yang benar. Selanjutnya pada tahapan kedua, ASC akan memilih *list* dari kata yang dianggap benar berdasarkan dengan kamus kata yang benar. Lalu pada tahapan terakhir ASC akan melakukan *ranking* pada kata yang dianggap benar dengan berdasarkan nilai probabilitas dari kesamaan (*similarity*) antar dua kata.

2.2.5 Text Mining

Istilah *text mining* mengacu pada proses dalam mencari dan menemukan pola unik serta pengetahuan yang penting pada sebuah dokumen teks yang tidak terstruktur (Kobayashi et al., 2018). *Text mining* memiliki beberapa pengaplikasian seperti *information extraction, text categorization,* dan *text summarization*. Pada *text categorization* merupakan teknik dalam mengkategorisasikan teks ke dalam suatu class yang telah didefinisikan sebelumnya (Mirończuk & Protasiewicz, 2018). *Class* merupakan suatu nilai yang berupa fakta yang dimiliki oleh teks tersebut, sehingga tujuan dari klasifikasi adalah untuk melakukan kategorisasi teks kedalam sebuah *class* dengan cara menemukan pola yang ada pada teks tersebut. Umumnya *text mining* memiliki lima tahapan dasar dalam melakukan pemrosesan pada sebuah dokumen teks yaitu *data collection, data transformation, identify data, analyze data, extract information* (Dang & Ahmad, 2014).

1. *Data Collection*

   Tahapan pertama merupakan tahapan untuk mendapatkan data yang tidak terstruktur beserta informasi yang ada didalamnya.

2. *Data Transformation*

   Tahapan kedua merupakan tahapan dalam mengubah data yang awalnya tidak terstruktur menjadi data yang terstruktur. Sebuah data tidak terstruktur akan sulit untuk dilakukan identifikasi dan analisis dikarenakan strukturnya yang tidak beraturan. Umumnya untuk melakukan tahapan tersebut diperlukan proses sebelumnya seperti *text preprocessing*.

3. *Identify Data*

   Tahapan ketiga merupakan tahapan untuk mengidentifikasi atau mengenali pola dari data yang sudah terstruktur pada tahapan sebelumnya.

4. *Analyze Data*

   Tahapan keempat merupakan tahapan untuk menganalisis pola dari data tersebut. Analisis pola bertujuan untuk mengetahui lebih detail kesesuaian terkait pola yang ditemukan dengan masalah yang ingin diselesaikan.

5. *Extract Information*

   Tahapan terakhir yaitu apabila pola telah dianalisis maka selanjutnya adalah mengeluarkan informasi yang berguna terkait data tersebut lalu menyimpannya.

2.2.6 Text Preprocessing

Text preprocessing merupakan metode dalam membersihkan data teks dengan mengidentifikasi dan mengeliminasi kata yang tidak diperlukan dalam proses klasifikasi (Lourdusamy & Abraham, 2018). Sehingga *text preprocessing* merupakan salah satu langkah penting yang memiliki dampak signifikan dalam proses klasifikasi pada *text mining*. Adapun tiga tahapan umum dalam melakukan *preprocessing* pada teks yaitu meliputi *tokenization*, *stopword removal*, dan *stemming* (Kadhim, 2018).

1. *Tokenization*

   Tokenization merupakan proses dalam membagi *string* pada konten tekstual sebagai kata, istilah, simbol, atau makna elemen lain menjadi sebuah token atau unit terkecil. Proses tokenisasi pada aslinya bukanlah tahapan pertama yang perlu dilakukan dalam melakukan *preprocessing* teks, namun terdapat

proses lain seperti melakukan *filtering tags hypertext markup language* (html), tanda baca, *case folding*, dll

2. *Stopword Removal*

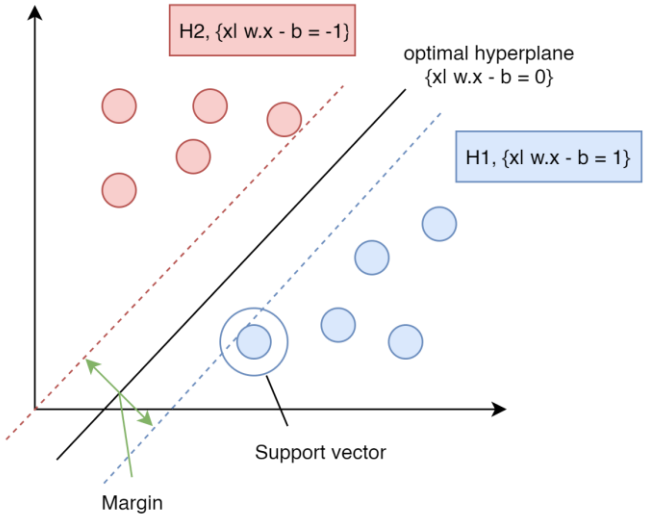
   Makna dari *stopword* merupakan kumpulan kata yang sering ditemui pada dokumen teks tanpa adanya keterkaitan dengan suatu topik tertentu pada dokumen teks tersebut. Sehingga dapat dikatakan bahwa *stopword* dapat dihilangkan saat melakukan klasifikasi dikarenakan tidak relevan apabila tetap digunakan.

3. *Stemming*

   Mengubah kata ke bentuk dasarnya merupakan istilah dari *stemming*. Adapun metode lain yang mirip dengan *stemming* yaitu *lemmatization*. Perbedaan mendasar dari keduanya adalah pada *lemmatization* mempedulikan konteks dari kata tersebut serta memperhatikan kata berdasarkan *Part of Speech (POS) tags*-nya namun pada *stemming* hal tersebut tidak dipedulikan.

2.2.7 Support Vector Machine

Support vector machine (SVM) merupakan salah satu algoritma *machine learning* untuk klasifikasi yang memiliki konsep yaitu mencari garis pemisah (*hyperplane*) yang terbaik yang dapat memisahkan antara dua buah *class* dari *input space* dengan memaksimalkan nilai pada margin (Rizwan et al., 2021). *Hyperplane* pada model SVM merupakan fungsi pemisah berbentuk model linear yang dapat membentuk *decision boundary* sebagai batas untuk memisahkan dua buah *class* (Chandra & Bedi, 2021). Model SVM dengan *hyperplane* terbaik adalah yang memiliki nilai margin paling besar (Isabelle et al., 2002). Margin merupakan istilah dari jarak antara titik data *point* yang paling dekat dengan *hyperplane* pada tiap *class*-nya. Titik data *point* yang paling dekat inilah yang disebut dengan *support vector*. Hal tersebut diasumsikan apabila data dapat dipisahkan secara linear (*linearly separable*). Terdapat ilustrasi Gambar 2.1 berikut untuk model SVM pada permasalahan *linearly separable*.



Gambar 2.3 Klasifikasi Linear Dua Kelas

(Mak & Montreal, 2000)

Dimisalkan adanya dua buah *class* yaitu -1 dan $+1$ dengan data yang tersedia dinotasikan sebagai $x_i \in R^d$ sedangkan pada label masing-masing dinotasikan $y_i \in \{-1, +1\}$ untuk $i = 1,2,\ldots, l$, yang mana l merupakan banyaknya data yang ada. Maka dua buah *class* tersebut dapat terpisah dengan benar oleh *hyperplane* yang memiliki dimensi d yang didefinisikan:

$$\overrightarrow{w} \cdot \vec{x} + b = 0 \quad (1)$$

Dimana w merupakan n-dimensional vektor dan b adalah *bias term*. Pola data *training* (x_i, y_i),....,(x_i, y_i) diberikan ketika x_i merupakan vektor berdimensi d dan $y_i = +1$ jika x_i dalam class A serta $y_i = -1$ jika x_i dalam class B. Pada data input yaitu x_i yang merupakan class $+1$ dapat dibuat perumusan dengan pertidaksamaan yaitu:

$$\overrightarrow{w} \cdot \vec{x} - b \geq 1 \text{ jika } y_i = +1 \quad (2)$$

Selanjutnya pada data input yaitu x_i yang merupakan class -1 juga dapat dibuat perumusan dengan pertidaksamaan yaitu:

$$\overrightarrow{w} \cdot \vec{x} - b \leq 1 \text{ jika } y_i = -1 \quad (3)$$

Apabila kedua pertidaksamaan tersebut digabungkan maka dapat menjadi pertidaksamaan sebagai berikut:

$$y_i(w \cdot x_i - b) \geq 1 \quad (4)$$

Telah dijelaskan juga bahwa SVM mencari *hyperplane* terbaik (*optimum hyperplane*) dengan cara memaksimalkan nilai margin. Sehingga untuk dapat memaksimalkan nilai margin, jarak (d) margin harus diukur dan dimaksimalkan menggunakan persamaan berikut:

$$d(w,b;x) = \frac{|(w^T x + b - 1) - (w^T x + b + 1)|}{||w||} = \frac{2}{||w||} \quad (5)$$

Penjelasannya yatu diasumsikan terdapat dua buah *hyperplane* bernilai $H1$ dengan persamaan $(w^T x + b - 1)$ yang setara dengan $\frac{|1+b|}{||w||}$ dan $H2$ dengan persamaan $(w^T x + b + 1)$ yang setara dengan $\frac{|b-1|}{||w||}$, kedua *hyperplane* tersebut membentuk posisi sejajar dengan *hyperplane* utama sehingga dapat disebut dengan *hyperplane support*. *Hyperplane support* tersebut memiliki margin yang memisahkan keduanya dengan persamaan $\frac{2}{||w||}$. Untuk dapat memaksimalkan $\frac{2}{||w||}$ yang mana setara dengan meminimalkan-nya $\frac{||w||^2}{2}$ yang artinya apabila menemukan $||w||$ terkecil diantara seluruh *hyperplane* yang ada pada batasannya, maka sebuah *hyperplane* $||w||$ yang memiliki nilai paling kecil itulah dapat dikatakan sebagai *optimum hyperplane*.

2.2.8 Feature Extraction

Feature extraction merupakan proses dalam mentransformasi *feature* yang ada pada teks menjadi struktur data kuantitatif seperti angka untuk dimasukan sebagai input pada proses pelatihan model (Zheng, Yoon, & Lam, 2014). Dalam *semantic feature*, *feature extraction* tersebut merupakan proses dalam mengekstrak *feature* yang berperan untuk mengetahui makna kata pada suatu teks. Salah satu metodenya adalah *term frequency-inverse document frequency* (TF-IDF). Sebagai contoh diberikan sebuah koleksi N dokumen, dimana f_{ij} merupakan banyaknya kemunculan suatu kata i di dalam dokumen j. sehingga TF dinotasikan dengan rumus:

$$TF_{ij} = \frac{f_{ij}}{max_k\, f_{kj}} \quad (6)$$

Dimana

f_{ij} : Banyaknya kata i pada dokumen j

$max_k\ f_{kj}$: Nilai maksimum k dari kemunculan kata apapun di dokumen j

Apabila diperlukan normalisasi TF-IDF maka pembagian nilai maksimum tersebut harus dilakukan (Moussa & Măndoiu, 2018). Pada IDF kata i dinotasikan dengan rumus:

$$IDF_i = log \frac{N}{n_i} \quad (7)$$

Dimana

N : Jumlah dokumen pada *corpus*

N_i : Banyaknya dokumen yang mengandung kata i pada N dokumen pada *corpus*

Sehingga rumus dari TF-IDF adalah:

$$\text{TF-IDF} = TF_{ij} \times IDF_I \quad (8)$$

Kata dengan nilai TF-IDF tertinggi memiliki makna yaitu kata tersebut muncul di banyak tempat pada satu dokumen, tetapi kemunculannya sedikit pada dokumen lain yang mengakibatkan kata tersebut memiliki nilai unik. Selanjutnya *feature extraction* pada *structural feature* merupakan proses dalam mengekstrak *feature* yang berperan untuk mengetahui struktur dan format pada sebuah dokumen teks. Beberapa implementasi dari *structural feature* adalah seperti perhitungan jumlah kata, jumlah kalimat, panjang ulasan, dan jumlah karakter.

2.2.9 Confusion Matrix

Confusion matrix merupakan tabel yang berfungsi untuk melakukan evaluasi dari hasil prediksi model yang terdiri dari *class* hasil prediksi dan *class* sebenarnya (Markoulidakis et al., 2021). Pada *confusion matrix*, kolom merepresentasikan *instance* dari *class* sebenarnya sementara tiap baris merepresentasikan *class* yang diprediksi. Apabila data yang ingin dilakukan klasifikasi memiliki dua *class* (*binary classification*) maka tabel *confusion matrix* memiliki dua *class* yaitu *class* positif dan *class* negatif. Adapun empat label yang dibentuk dari dua *class* tersebut pada *confusion matrix* yaitu: *True Positive* (TP),

False Negative (FN), *True Negative* (TN), dan *False Positive* (FP) seperti yang ada pada Tabel 2.2 berikut.

Tabel 2.2 Matriks *Confusion*

Matriks *Confusion*		*Actual Value*	
		Positive	*Negative*
Predicted Value	*Positive*	TP	FP
	Negative	FN	TN

Label matriks *confusion* tersebut bermaksud untuk membedakan antara hasil prediksi model dengan *class* sebenarnya (*actual value*) dengan penjelasan masing-masing yaitu label *True Positive* (TP) merupakan jumlah data yang terprediksi positif dan benar positif. Lalu pada label *False Negative* (FN) merupakan jumlah data yang terprediksi negatif namun positif. Sedangkan label *False Positive* (FP) merupakan jumlah data yang terprediksi positif namun negatif. Terakhir yaitu label *True Negative* (TN) merupakan jumlah data yang terprediksi negatif dan benar negatif.

Dari label yang ada pada *confusion matrix*, didapatkan matriks evaluasi lainnya untuk mengukur keberhasilan klasifikasi model yaitu *accuracy*, *precision*, *recall*, dan *f-measure*. Penjelasan dari tiap matriks evaluasinya adalah sebagai berikut:

1. *Accuracy*, merupakan nilai dari pembagian seluruh data yang di klasifikasikan oleh model bernilai benar dengan jumlah seluruh *dataset*. Sehingga semakin banyak data yang di klasifikasikan bernilai benar maka semakin tinggi nilai akurasinya. Berikut merupakan rumus dari akurasi.

$$Accuracy = \frac{(TP+TN)}{(TP+FN+FP+TN)} \quad (9)$$

2. *Precision*, nilai yang didapatkan dari hasil pembagian antara benar positif dengan total hasil yang diprediksi bernilai positif. Sehingga *precision* merupakan penilaian model dalam memprediksi nilai positif dengan total nilai yang diprediksi positif (*positive predictive value*). Berikut rumusnya.

$$Precision = \frac{TP}{(TP+FP)} \quad (10)$$

3. *Recall*, nilai yang didapatkan dari hasil pembagian antara benar positif dengan keseluruhan yang seharusnya diprediksi bernilai positif. *Recall* juga disebut sensitivitas. Berikut rumusnya.

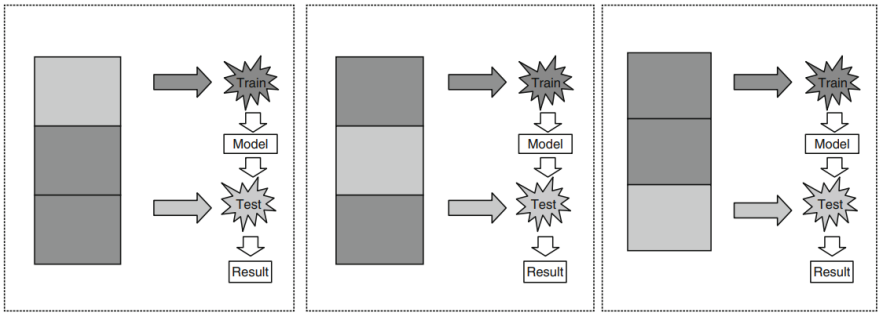
$$Recall = \frac{TP}{(TP+FN)} \quad (11)$$

4. *F-measure* merupakan nilai penyeimbang (*harmonic average*) antara *precision* dan *recall*. Semakin tinggi nilai *f-measure* maka semakin tinggi efisiensi dari model dimana nilai satu merupakan nilai terbaik sementara nol merupakan nilai terburuk. Berikut rumusnya.

$$F\text{-}measure = 2 \times \frac{precision \star recall}{precision+recall} \quad (12)$$

2.2.10 K-fold Cross Validation

Cross validation merupakan metode untuk memberikan penilaian pada performa model klasifikasi saat melakukan klasifikasi pada data yang baru dengan melakukan proses iterasi yang melibatkan pembagian dua sampel data yaitu menjadi data latih dan data uji (Moreno-Torres, Saez, & Herrera, 2012). Salah satu jenis dari *cross validation* adalah *k-fold cross validation*, dimana pada metode ini pertama data akan dibagi menjadi sebanyak *k* bagian yang biasa disebut *segment* atau *folds*. Kemudian iterasi sebanyak *k* dilakukan sedemikian rupa sehingga dalam setiap iterasi terdapat *fold* (bagian) data yang berbeda yang digunakan sebagai data validasi (*test* data), sedangkan sisanya $(k - 1)$ bagian digunakan sebagai data latih (*training* data).



Gambar 2.4 Model *3-fold Cross Validation*

(Refaeilzadeh et al., 2020)

Gambar 2.4 merupakan implementasi dari *3-fold cross validation.* Pada bagian yang berwarna gelap yaitu $(k - 1)$ dengan jumlah dua digunakan sebagai data *training*, sementara bagian yang berwarna terang digunakan sebagai data *testing*. Diasumsikan nama dari tiap bagian pada data adalah $k1, k2$ dan $k3$ maka tahapan dari proses iterasinya adalah sebagai berikut:

1. Dikarenakan nilai dari k adalah tiga maka data akan dibagi sebanyak tiga bagian dan akan dilakukan iterasi sebanyak tiga kali.
2. Pada iterasi pertama data $k1$ digunakan sebagai data *testing* dan $k2, k3$ akan digunakan sebagai data *training*.
3. Iterasi kedua data $k2$ digunakan sebagai data *testing* dan $k1, k3$ akan digunakan sebagai data *training*.
4. Terakhir yaitu pada tahap keempat data $k3$ digunakan sebagai data *testing* dan $k1, k2$ akan digunakan sebagai data *training*.

Dari setiap iterasi tersebut akan dihitung dan dicatat akurasi dari performa model dalam melakukan klasifikasi, sehingga hasil akhirnya adalah nilai akumulasi dari total akurasi yang didapat pada tiap iterasi tersebut.

2.3 Integrasi Keilmuan

Ulasan merupakan bentuk tanggapan yang diberikan oleh seseorang yang menyatakan pendapat mereka terkait suatu hal. Hasil wawancara dengan Bapak Bahtiyar Rifa'i yang merupakan koordinator dari pusat ma'had al-jamiah Universitas Islam Negeri Sunan Ampel serta dosen luar biasa untuk mata kuliah dasar umum studi hadits dan studi Al-quran menjelaskan bahwa terdapat keterkaitan antara ulasan yang diberikan dengan kebermanfaatan yang dibawa oleh ulasan tersebut sehingga menjadikannya ulasan yang bermanfaat. Dimana hal tersebut adalah hal yang ingin penelitian ini capai yaitu mengetahui ulasan mana yang bermanfaat atau tidak bermanfaat. Seperti pada Firman Allah SWT dalam surah Al-Ma'idah ayat 8 yang berbunyi,

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْا كُوْنُوْا قَوّٰمِيْنَ لِلّٰهِ شُهَدَاۤءَ بِالْقِسْطِۖ وَلَا يَجْرِمَنَّكُمْ شَنَاٰنُ قَوْمٍ عَلٰٓى
اَلَّا تَعْدِلُوْا ۗاِعْدِلُوْاۗ هُوَ اَقْرَبُ لِلتَّقْوٰىۖ وَاتَّقُوا اللّٰهَ ۗاِنَّ اللّٰهَ خَبِيْرٌۢ بِمَا تَعْمَلُوْنَ

Artinya:

"Wahai orang-orang yang beriman! Jadilah kamu sebagai penegak keadilan karena Allah, (ketika) menjadi saksi dengan adil. Dan janganlah kebencianmu terhadap suatu kaum mendorong kamu untuk berlaku tidak adil. Berlaku adillah. Karena (adil) itu lebih dekat kepada takwa. Dan bertakwalah kepada Allah, sungguh, Allah Maha Teliti terhadap apa yang kamu kerjakan". (QS. Al-Ma'idah: 8)

Ayat tersebut menjelaskan bahwa Allah SWT menuntun agar umat muslim selalu berlaku adil yaitu menegakkan kebenaran kepada siapapun meski kepada orang-orang yang tidak disukai. Seperti ketika memberikan sebuah ulasan, jangan sampai dikarenakan terdapat kebencian terhadap suatu kaum tertentu menjadikan ulasan tersebut tidak disampaikan secara benar dan terkesan asal-asalan sehingga dapat mengurangi nilai kebermanfaatan dari ulasan tersebut. Dikarenakan ulasan memiliki manfaat (bermanfaat) apabila disampaikan secara benar dan tidak asal-asalan. Ketika menyampaikan sesuatu dengan benar dan tidak asal-asalan, hendaklah juga diikuti dengan menyampaikannya secara baik. Hal tersebut dijelaskan pada surah Muhammad ayat 21 yang berbunyi,

طَاعَةٌ وَّقَوْلٌ مَّعْرُوْفٌۗ فَاِذَا عَزَمَ الْاَمْرُۗ فَلَوْ صَدَقُوا اللّٰهَ لَكَانَ خَيْرًا لَّهُمْۚ

Artinya:

(Yang lebih baik bagi mereka adalah) taat (kepada Allah) dan bertutur kata yang baik. Sebab apabila perintah (perang) ditetapkan (mereka tidak menyukainya). Padahal jika mereka benar-benar (beriman) kepada Allah, niscaya yang demikian itu lebih baik bagi mereka. (QS. Muhammad: 21)

Ayat tersebut menjelaskan bahwa ketika memberikan suatu ulasan maka katakan-lah dengan baik agar tidak ada orang lain yang tersinggung atas tulisan maupun ucapan yang telah diberikan, yang mana hal tersebut juga membuktikan bahwa orang yang memberikan ulasan benar-benar beriman kepada Allah SWT dan itu akan bernilai lebih baik juga bagi orang tersebut. Ketika memberikan ulasan

juga harus diikuti dengan berkata jujur yaitu mengatakan dengan sebenar-benarnya, hal tersebut sesuai dengan surah Al-Ahzab ayat 70-71 yang berbunyi,

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوا اتَّقُوا اللّٰهَ وَقُوْلُوْا قَوْلًا سَدِيْدًاۙ (٧٠) يُّصْلِحْ لَكُمْ اَعْمَالَكُمْ وَيَغْفِرْ لَكُمْ ذُنُوْبَكُمْۗ وَمَنْ يُّطِعِ اللّٰهَ وَرَسُوْلَه فَقَدْ فَازَ فَوْزًا عَظِيْمًا (٧١)

Artinya:

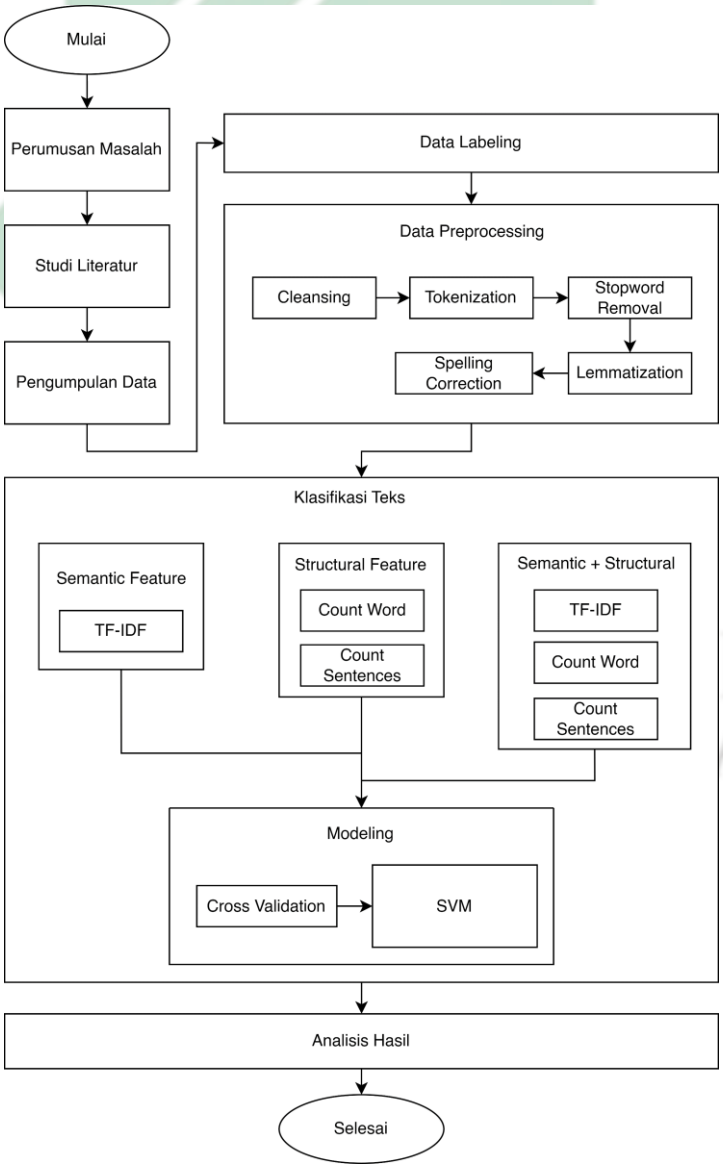
Wahai orang-orang yang beriman! Bertakwalah kamu kepada Allah dan ucapkanlah perkataan yang benar, niscaya Allah akan memperbaiki amal-amalmu dan mengampuni dosa-dosamu. Dan barangsiapa menaati Allah dan Rasul-Nya, maka sungguh, dia menang dengan kemenangan yang agung. (QS. Al-Ahzab: 70-71)

Ayat tersebut menjelaskan bahwa ketika memberikan suatu ulasan maka katakanlah dengan jujur yang mana Allah SWT akan memperbaiki amalan ibadah dan mengampuni segala dosa-dosanya bagi mereka yang berkata jujur. Dari hal tersebut maka dengan menerapkan suatu metode pada penelitian ini, diharapkan dapat membantu dalam mengidentifikasi ulasan mana yang bermanfaat atau tidak sehingga dapat mempermudah dalam menilai kebermanfaatan dari isi ulasan tersebut.

BAB III
METODOLOGI PENELITIAN

3.1 Tahapan Penelitian

Tahapan yang akan dilakukan dalam penelitian ini direpresentasikan kedalam ilustrasi berupa diagram alur. Hal tersebut bertujuan untuk mempermudah penyusunan penelitian serta menyampaikan dan memahami secara rinci terkait langkah-langkah dari penelitian yang akan dilakukan. Ilustrasi dari diagram alur penelitian dapat dilihat pada Gambar 3.1 berikut.



Gambar 3.1 Diagram Alur Penelitian

Berdasarkan Gambar 3.1 terdapat tujuh tahapan penelitian yang harus dilakukan untuk menyelesaikan penelitian klasifikasi kualitas ulasan produk berdasarkan *semantic* dan *structural feature*. Berikut merupakan penjelasan detail mengenai tiap langkah yang akan dilakukan pada penelitian ini.

3.1.1 Perumusan Masalah

Dalam melakukan sebuah penelitian, perumusan masalah diperlukan sebagai landasan dalam melakukan penelitian tersebut. Masalah yang dirumuskan pada penelitian ini telah tertera pada latar belakang yaitu mengenai klasifikasi kualitas ulasan produk berdasarkan *semantic* dan *structural feature* menggunakan model SVM. Klasifikasi tersebut bertujuan untuk memprediksi ulasan mana yang bermanfaat atau tidak.

3.1.2 Studi Literatur

Pada studi literatur dilakukannya pembelajaran terhadap literatur terdahulu dalam memahami konsep dan variabel yang relevan yang akan digunakan pada penelitian ini. Literatur tersebut didapatkan dari penelitian yang sudah dilakukan sebelumnya seperti mengenai implementasi model SVM, ekstraksi ciri pada teks menggunakan *semantic* dan *structural feature*, serta klasifikasi teks ulasan. Sumber untuk studi literatur lainnya yang digunakan yaitu jurnal dan buku.

3.1.3 Pengumpulan Data

Dalam penelitian ini data yang digunakan adalah *dataset* ulasan produk dari amazon tahun 2015. *Dataset* yang digunakan merupakan *dataset* dengan jenis terbuka (*open dataset*) yang didapatkan dari situs resminya di "https://s3.amazonaws.com/amazon-reviews-pds/readme.html". Adapun *dataset* yang digunakan merupakan hasil pengumpulan data ulasan produk dari berbagai kategori dari tahun 1995 sampai 2015. Terdapat 15 variabel yang ada, namun pada penelitian ini dibatasi hanya digunakan enam variabel. Berikut pada Tabel 3.1 merupakan variabel yang digunakan beserta nilai dan keterangannya.

Tabel 3.1 Variabel Data Penelitian

No.	Variabel	Nilai	Keterangan
1.	*customer_id*	Sejumlah nilai angka *random* (contoh: 302120, 445)	*Identifier* yang bernilai *random* yang dapat digunakan untuk meng-agregasi ulasan yang ditulis oleh satu customer
2.	*product_id*	Sejumlah nilai kombinasi angka dan huruf (contoh: B00MUTIDKI, B001AMHWP8)	*Unique id* dari suatu produk
3.	*helpful_votes*	Sejumlah nilai angka (contoh: 12, 3, 1)	Total *vote* manfaat yang diterima oleh suatu ulasan
4.	*total_votes*	Sejumlah nilai angka (contoh: 34, 12, 7)	Total *vote* baik *vote* manfaat (*like*) maupun *vote* tidak manfaat (*dislike*) yang diterima oleh suatu ulasan
5.	*verified_ purchase*	Sejumlah huruf bernilai 'Y' atau 'N'	Bernilai 'Y' apabila customer telah terverifikasi membeli produk langsung di amazon dan tanpa diberi diskon berlebih. Sedangkan bernilai 'N' apabila *customer* membeli produk tidak langsung melalui amazon dan tidak membayar dengan harga yang tersedia untuk sebagaian besar pembeli di amazon
6.	*review_body*	Sejumlah kata	Isi dari ulasan yang diberikan

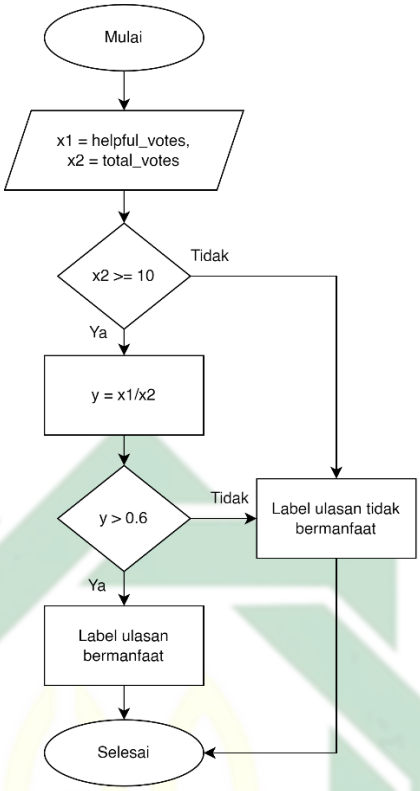
Dataset yang digunakan pada penelitian ini akan dibagi menjadi dua kategori yaitu *search product* dan *experience product*. *Search product* merupakan kategori produk yang konsumen dapat nilai sebelum membeli produk tersebut. Sedangkan pada *experience product* merupakan produk yang konsumen tidak dapat nilai sebelum membeli produk tersebut sehingga harus dicoba terlebih dahulu. Pada *search product* akan digunakan *dataset video games* sedangkan *experience product* akan digunakan *dataset* produk kecantikan.

3.1.4 Data Labeling

Pada tahapan *data labeling* ulasan akan diberikan sebuah label. Label tersebut menunjukan bahwa sebuah ulasan masuk kedalam ulasan yang bermanfaat atau tidak bermanfaat. Berdasarkan penelitian Ghose & Ipeirotis (2011) yang juga menggunakan *dataset product review* amazon, sebuah ulasan dapat dikatakan bermanfaat apabila nilai rasio manfaat yang ada pada ulasan tersebut adalah bernilai lebih dari 0.6. Dimana nilai rasio manfaat didapatkan dari pembagian antara $helpful\_votes$ dengan $total\_votes$ (Korfiatis, García-Bariocanal, & Sánchez-Alonso, 2012).

Dalam menentukan nilai 0.6 sebagai batasan (*threshold*) untuk meng-klasifikasikan ulasan tersebut masuk kedalam ulasan yang bermanfaat atau tidak bermanfaat. Penelitian Ghose & Ipeirotis pertama mengasumsikan bahwa *helpfulness* dari suatu ulasan bernilai antara nol sampai dengan satu (nilai rasio manfaat). Dimana hal tersebut maka, *helpfulness* ulasan dapat diberikan suatu batasan. Untuk dapat memilih batasan yang baik yaitu yang dapat meng-klasifikasikan ulasan masuk kedalam ulasan bermanfaat atau tidak, Ghose & Ipeirotis menggunakan analisis Receiver Operator Characteristic (ROC) dengan membandingkannya pada hasil pelabelan 1000 ulasan secara acak yang dilakukan manual oleh dua pakar. Dimana dengan menggunakan *kappa statistic* hasil pelabelan manual tersebut menunjukan *substantial agreement* dengan nilai $k = 0.739$. Selanjutnya pada hasil membandingkannya menunjukan bahwa apabila ditetapkan batasan bernilai 0.6 maka *error rates* dari analisis ROC dapat diminimalkan yang menunjukan bahwa apabila ulasan memiliki nilai rasio manfaat lebih dari 0.6 maka ulasan tersebut dapat dilabelkan sebagai ulasan bermanfaat, begitu juga dengan sebaliknya.

Nilai 0.6 ini selanjutnya digunakan pada banyak penelitian yang sama yaitu yang menggunakan *dataset product review* amazon serta melakukan klasifikasi kualitas ulasan produk seperti penelitian Krishnamoorthy (2015) dan M. S.I. Malik & Hussain (2017). Sehingga pada penelitian ini juga akan digunakan nilai 0.6 tersebut. Selanjutnya dari semua hal tersebut maka dapat ditentukan alur dari *data labeling* yaitu sebagaimana Gambar 3.2 berikut.



Gambar 3.2 Alur *Data Labeling*

Adapun ulasan yang memiliki total *vote* kurang dari 10 dapat dikatakan bahwa ulasan tersebut tidak bermanfaat dikarenakan sedikitnya konsumen lain yang memberikan penilaian pada ulasan tersebut. Hal itu juga untuk memastikan bahwa proses klasifikasi memiliki data yang representatif atau dalam kata lain data tersebut benar-benar mewakili ulasan yang bermanfaat atau tidak bermanfaat. Berikut pada Tabel 3.2 merupakan contoh dari *data labeling.*

Tabel 3.2 Contoh *Data Labeling*

helpful_votes	*total_votes*	*helpful_ratio*	*review_body*	Label
2	4	0.5	This was not what I expected.	Tidak Bermanfaat
8	8	1	Did not help me I feel it makes it worse still have bags under my eyes.	Tidak Bermanfaat
15	15	1	Bought as a gift for a friend- she loves it, and wears it often. Nice quality.	Bermanfaat

6	11	0.54	We use these all of the time, and they really do help to get some of the ick off when you can't get to a sink.	Tidak Bermanfaat

3.1.5 Data Preprocessing

Setelah *data labeling* maka tahapan berikutnya adalah melakukan *preprocessing* pada data ulasan yang sudah diberi label. *Preprocessing* tersebut bertujuan untuk mendapatkan data yang bersih. Terdapat enam proses yang akan dilakukan yang dijelaskan sebagai berikut:

1. *Data Cleansing*

Pada proses *data cleansing* akan dibagi menjadi dua tahapan yaitu *cleansing satu* dan *cleansing dua*. Tahapan *cleansing satu* merupakan tahapan untuk mendapatkan data ulasan yang dapat diandalkan (valid). Beberapa prosesnya yaitu pembersihan pada data ulasan yang kurang lengkap/kosong, memastikan bahwa konsumen yang memberikan ulasan telah terverifikasi membeli produk tersebut, dan melakukan penghapusan pada data yang duplikat. Berikut merupakan *pseudocode* yang diterapkan:

Cleansing satu

Deklarasi:

```
var data_ulasan
```

```
START

For i in data_ulasan:
   If empty(data_ulasan) == True
      Then remove data_ulasan

   If data_ulasan.verified_purchase == 'N'
      Then remove data_ulasan

   If duplicate(data_ulasan.product_id and
                data_ulasan.customer_id)
      Then remove data_ulasan

END for
OUTPUT validated_data_ulasan

END
```

Setelah mendapatkan data ulasan yang valid maka tahapan *cleansing* selanjutnya adalah untuk mendapatkan data ulasan yang bersih secara strukturnya. Beberapa prosesnya yaitu mengubah isi ulasan menjadi teks berukuran normal (*text normalization*), menghilangkan *uniform resource locator (url)* dan *html tags*, memperluas kata singkatan dan kata gaul, serta menghilangkan tanda baca, simbol, dan angka. Hal tersebut sebagaimana *pseudocode* berikut:

Cleansing dua

```
Deklarasi:
var validated_data_ulasan

START

For i in validated_data_ulasan:
   Transform to lowercase
   Remove url
   Remove html tags
   Expand contraction word, slang word
   Remove punctuation, symbol, number

END for
OUTPUT clean_data_ulasan

END
```

2. *Tokenization*

Tokenization digunakan untuk memisahkan kalimat menjadi token kata. Token kata hasil *tokenization* tersebut nantinya akan berguna pada proses berikutnya seperti *stopword removal*. Proses *tokenization* dilakukan dengan cara mengidentifikasi *whitespace* pada setiap kata sehingga dapat diubah menjadi sebuah token kata.

3. *Stopword Removal*

Stopword removal merupakan proses dalam menghapus kata umum yang ada pada bahasa tersebut. Kata umum (*stopword*) tersebut umumnya tidak terikat pada suatu topik tertentu sehingga tidak memiliki makna dan perlu dihilangkan. List dari *stopword* yang akan dihilangkan dalam penelitian ini diambil dari *library Natural Language Toolkit* (NLTK). Bahasa yang digunakan pada pengaturan *library* adalah bahasa inggris.

4. *Lemmatization*

Lemmatization merupakan proses dalam mengubah kata ke bentuk dasarnya namun dengan tetap memperhatikan isi serta makna dari kata tersebut. Meskipun tidak ada perbedaan yang banyak setelah tahapan *stopword removal*, *lemmatization* tetap perlu dilakukan untuk memastikan bahwa kata tersebut telah diubah ke bentuk dasarnya. *Lemmatization* akan diterapkan dengan bantuan *POS tagging* dari *library* NLTK.

5. *Spelling Correction*

Spelling correction merupakan proses dalam melakukan koreksi pada kata yang salah dalam ejaannya. Pada proses *spelling correction* suatu kata dilakukan cek terlebih dahulu yaitu apakah ada atau tidak di dalam kamus kata bahasa inggris (*WordNet*). Selanjutnya apabila kata tersebut tidak ada di *WordNet* maka kata tersebut memiliki kemungkinan terdapat kesalahan dalam ejaannya yang mana perlu diperbaiki. Dalam proses perbaikan ejaan kata-nya akan digunakan *library TextBlob*. Konsep perbaikannya sesuai dengan Gambar 2.2 pada bab dua.

3.1.6 Klasifikasi Teks

Pada tahapan klasifikasi akan dilakukan perhitungan pada *dataset* ulasan produk yang telah diberi label dan di *preprocessing* sebelumnya. Adapun beberapa prosesnya sebagai berikut:

1. *Feature Extraction*

Proses pertama adalah melakukan ekstraksi pada masing-masing fiturnya yaitu *semantic* dan *structural features*. Proses ini akan mengubah data yang sebelumnya berupa teks menjadi data berupa angka. Pada *semantic feature* proses ekstraksi akan dilakukan menggunakan TF-IDF dimana data teks yang digunakan merupakan data teks yang telah dilakukan *preprocessing* sebelumnya. Pada Tabel 3.3 merupakan contoh hasil dari perhitungan TF-IDF sebagaimana rumus persamaan 8.

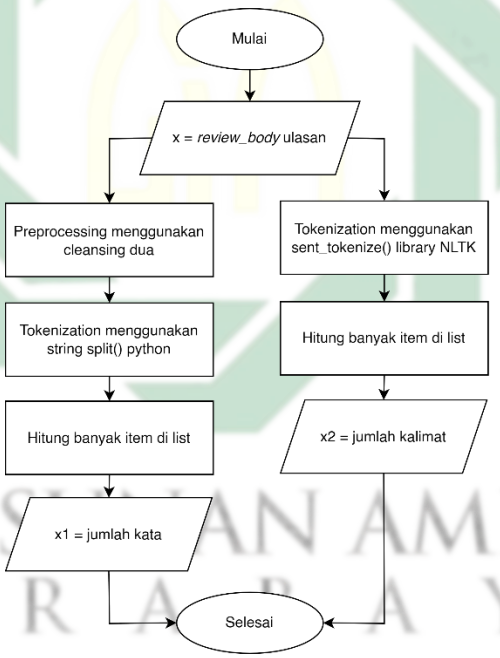
Tabel 3.3 Contoh Hasil Pembobotan TF-IDF

Dokumen ke-	Kata ke-				Label
	0	1	2	...	
0	0.000000	0.469791	0.580286	...	Tidak Bermanfaat
1	0.000000	0.687624	0.000000	...	Tidak Bermanfaat

2	0.511849	0.000000	0.000000	...	Bermanfaat
3	0.000000	0.469791	0.580286	...	Tidak Bermanfaat

Pada *semantic feature*, hasil dari TF-IDF akan memiliki jumlah kolom *feature* sebanyak perhitungan n kata pada j dokumen. Dimana semakin banyak kata yang diberikan bobot oleh TF-IDF pada tiap dokumennya maka jumlah kolom *feature* juga akan semakin banyak. Hasil TF-IDF tersebut nantinya akan digunakan sebagai input pada data *training* dan data *testing* model SVM.

Selanjutnya pada *structural feature* akan dilakukan perhitungan struktur dari teks dengan cara menghitung jumlah kata dan jumlah kalimat pada ulasan. Perhitungan jumlah kata dan jumlah kalimat tersebut didapatkan dari variabel $review\_body$. *Flow* perhitungannya dapat dilihat pada Gambar 3.3 berikut.



Gambar 3.3 Alur Ekstraksi *Structural Feature*

Pada ektraksi fitur jumlah kata akan dilakukan perhitungan dengan cara melakukan *preprocessing* terlebih dahulu. *Preprocessing* tersebut bertujuan untuk menghilangkan karakter-karakter yang bukan kata seperti tanda baca atau simbol. Selanjutnya dilakukan *tokenization* menggunakan *method split* pada bahasa pemrogaman *python* untuk menghasilkan sebuah list yang menampung kata-kata dari ulasan. Setelahnya dihitung banyak item yang ada di list tersebut untuk mendapatkan jumlah kata.

Sedangkan pada ektraksi fitur jumlah kalimat akan dilakukan menggunakan *method sent_tokenize* pada *library* NLTK. *Method* tersebut akan menghasilkan list yang menampung kalimat-kalimat dari suatu ulasan. Setelahnya akan dihitung banyak item yang ada di list tersebut untuk mendapatkan jumlah kalimat.

Setelah didapatkan jumlah kata dan jumlah kalimat dari suatu ulasan maka nilai-nilai jumlah tersebut akan direpresentasikan menjadi dua kolom *features* yang digunakan sebagai data *training* dan data *testing* pada model SVM sebagaimana pada Tabel 3.4.

Tabel 3.4 Contoh Ekstraksi *Structural Feature*

Dokumen ke-	X1 (jumlah_kata)	X2 (jumlah_kalimat)	Label
0	34	3	Tidak Bermanfaat
1	56	7	Tidak Bermanfaat
2	45	5	Bermanfaat
3	125	12	Tidak Bermanfaat

2. Proses Klasifikasi

Untuk menentukan seberapa baik model dalam memprediksi suatu ulasan yaitu apakah ulasan tersebut bermanfaat atau tidak dengan menggunakan *features* yang telah diusulkan (*semantic* dan *structural*), maka pada proses klasifikasinya akan dibagi kedalam tiga skenario. Pembagian tiga skenario tersebut juga akan diujikan pada masing-masing dua kategori *dataset* yaitu *search product* (*video games*) dan *experience product* (produk kecantikan). Pada Tabel 3.5 merupakan gambaran dari ketiga skenario tersebut.

Tabel 3.5 Skenario Klasifikasi SVM

Skenario ke-	Nama *Feature*	Keterangan
1	*Semantic*	Skenario pertama *dataset* akan dilakukan klasifikasi berdasarkan *semantic feature* menggunakan TF-IDF
2	*Structural*	Pada skenario kedua *dataset* akan dilakukan klasifikasi berdasarkan *structural feature* menggunakan perhitungan sebelumnya
3	*Semantic* +	Skenario terakhir *dataset* akan dilakukan klasifikasi

	Structural	dengan menggabungkan kedua *features* tersebut menjadi satu

Dari beberapa skenario klasifikasi tersebut akan dibandingkan hasilnya antara satu dengan yang lain untuk mengetahui *feature* mana yang paling berpengaruh dalam memprediksi ulasan yang bermanfaat atau tidak bermanfaat. Adapun pada skenario terakhir adalah penggabungan dari kedua *features* yang ada untuk mengetahui apakah hasil penggabungan tersebut memiliki efek pada nilai akurasi klasifikasi. Pada proses penggabungannya, kedua *features* tersebut akan digabungkan menjadi satu sebagaimana contoh pada Tabel 3.6 berikut.

Tabel 3.6 Contoh Skenario Penggabungan Kedua *Features*

Dokumen ke-	Kata ke-				X1	X2	Label
	0	1	2	...			
0	0.000000	0.469791	0.580286	...	34	3	Tidak Bermanfaat
1	0.000000	0.687624	0.000000	...	56	7	Tidak Bermanfaat
2	0.511849	0.000000	0.000000	...	45	5	Bermanfaat
3	0.000000	0.469791	0.580286	...	125	12	Tidak Bermanfaat

Pada Tabel 3.6 di atas *semantic feature* berada dibagian kiri pada tabel, sedangkan untuk *structural feature* berada dibagian kanan tabel dengan diikuti oleh label yang akan diprediksi nilainya. Pada proses *training* dan *testing* nantinya, akan digunakan *row* dari tiap dokumen teksnya dimana *X* sebagai *feature* yaitu penggabungan antara fitur *semantic* dan *structural* tersebut dan *y* sebagai label yang akan diprediksi.

3. Evaluasi Model

Untuk mengevaluasi performa dari model pada proses ini akan digunakan matriks *confusion* yang didalamnya terdiri dari *accuracy*, *precision*, *recall*, dan *f-measure* dengan perhitungan sesuai pada rumus 9, rumus 10, rumus 11, dan rumus 12. Evaluasi ini juga mengukur bagaimana pengaruh penggunaan *feature* yang diusulkan dalam penelitian sehingga dapat mengetahui *feature* mana yang paling berpengaruh dalam melakukan klasifikasi ulasan bermanfaat. Adapun proses *training* dan *testing* model akan digunakan *k-fold cross validation* dengan nilai *k* adalah 10. Penggunaan *10-fold cross validation* bermaksud untuk melakukan

validasi pada hasil yang dikeluarkan oleh model sehingga dapat teruji dan diyakini kebenarannya.

3.1.7 Analisis Hasil

Tahapan ini akan dilakukan pembahasan terkait analisis pada hasil evaluasi dari model klasifikasi serta penerapan *features* yang diusulkan dalam penelitian ini. Yaitu apakah model dan *features* yang diusulkan dapat membantu dalam mengidentifikasi ulasan mana yang bermanfaat atau tidak bermanfaat. Setelahnya akan dibuat kesimpulan untuk menjawab rumusan masalah yang telah dibahas pada bab sebelumnya. Adapun nantinya terdapat saran sebagai masukan pada penelitian selanjutnya

BAB IV
HASIL DAN PEMBAHASAN

4.1 Pengumpulan Data

Proses awal dari penelitian ini adalah mendapatkan data yang akan digunakan. Data yang digunakan merupakan data sekunder berupa *dataset* ulasan produk kecantikan (*experience product*) dan ulasan produk video games (*search product*) dari amazon yang didapatkan dengan cara mengunduhnya secara langsung melalui situs resminya yaitu "https://s3.amazonaws.com/amazon-reviews-pds/readme.html". Terdapat 15 variabel yang ada pada *dataset* sebagaimana Tabel 4.1 berikut.

Tabel 4.1 Variabel *Dataset*

No.	Variabel	Tipe Data	Keterangan
1.	*marketplace*	*String*	Dua digit kode negara
2.	*customer_id*	*Number*	*Identifier customer*
3.	*review_id*	*String + number*	*Identifier* ulasan
4.	*product_id*	*String + number*	*Identifier* produk
5.	*product_parent*	*Number*	*Identifier* untuk mengagregasi ulasan dalam satu produk
6.	*product_title*	*String*	Judul produk
7.	*product_category*	*String*	Kategori produk
8.	*star_rating*	*Number*	Rating yang diberikan oleh ulasan
9.	*helpful_votes*	*Number*	Total *vote* manfaat (*like*) yang diterima oleh ulasan
10.	*total_votes*	*Number*	Total *vote like* dan *dislike* yang diterima oleh ulasan
11.	*vine*	*String*	Ulasan ditulis sebagai bagian dari program vine amazon
12.	*verified_purchase*	*String*	Ulasan terverifikasi telah membeli produk
13.	*review_headline*	*String*	Judul ulasan
14.	*review_body*	*String*	Isi ulasan
15.	*review_date*	*Date*	Tanggal ulasan ditulis

Ekstensi file dari *dataset* yang telah di download adalah berupa *tab-separated values* (.tsv). Untuk dapat dilakukannya analisis maka ekstensi file tersebut perlu diubah terlebih dahulu menjadi *comma-seperated values* (.csv).

Setelah dilakukan pengubahan pada *dataset* maka hasil dari pengumpulan data adalah sebagaimana Tabel 4.2 berikut.

Tabel 4.2 Hasil Pengumpulan Data

marketplace	*customer_id*	*review_id*	...	*review_headline*	*review_body*	*review_date*
US	27433333	RFGPW2...	...	Four Stars	High price...	8/24/2015
US	44655780	R2J7DA...	...	beautiful!	Beautiful!...	8/24/2015
US	51252428	R31HWC...	...	Smells so good.	Can't keep...	8/24/2015
US	44964995	R1AGJS...	...	Five Stars	I am very...	8/24/2015
US	18983217	RO5QQ...	...	Repurchase	My skin h...	8/24/2015
...	...	...	...	...	...	...
US	22904591	R3TQM...	...	Don't bother...	These bare...	8/24/2015
US	18625244	R35CIS4...	...	Well worth it!!..	These are...	8/24/2015
US	18258673	R1BZVD...	...	Low quality, ...	Worked gr...	8/24/2015
US	22904591	RLPW0S...	...	My favorite c...	My favorit...	8/24/2015
US	116326	RYVSQJ...	...	I love this stuf...	Absolutely...	8/24/2015

Hasil pengumpulan data tersebut berlaku pada kedua *dataset* yaitu ulasan produk kecantikan dan ulasan produk *video games*. Seperti pada jumlah keseluruhan data yang telah didapatkan adalah sejumlah 1.048.576 baris. Serta juga terdapat 15 variabel yang ada dengan nilai sedemikan rupa menyesuaikan dengan *dataset* yang digunakan.

4.2 Pelabelan Data Ulasan

Ulasan yang dapat membantu calon konsumen dalam memutuskan pembelian dari suatu produk dapat diidentifikasi dengan cara melihat *helpful votes* yang diterima oleh ulasan tersebut. Sebagaimana pada metodologi penelitian maka hasil dari pelabelan ulasan untuk masing-masing *dataset* adalah pada Tabel 4.3 dan Tabel 4.4 berikut.

Tabel 4.3 Hasil Pelabelan Data Ulasan Produk Kecantikan

helpful_votes	*total_votes*	...	*review_body*	*review_date*	*label*
10	27	...	Looking at this product...	5/15/2014	tidak bermanfaat
6	8	...	I have been using this s...	2/13/2015	tidak bermanfaat
...	...	...	...	...	...
18	23	...	First off, i'll say I'm sk...	5/15/2014	bermanfaat

12	12	...	I have been using the al...	1/20/2015	bermanfaat

Hasil pelabelan dari tiap ulasan pada *dataset* adalah dengan cara melihat $helpful\_votes$ dan $total\_votes$ yang diterima oleh ulasan tersebut. Suatu ulasan dapat dilabelkan bermanfaat apabila nilai bagi dari $helpful\_votes$ dengan $total\_votes$ adalah bernilai minimal 0.6 dan nilai $total\_votes$ tidak kurang dari 10. Sebagaimana pada Tabel 4.3 merupakan contoh hasil pelabelan pada *dataset* produk kecantikan yang mana hal tersebut juga berlaku pada *dataset* produk *video games* seperti pada Tabel 4.4 berikut.

Tabel 4.4 Hasil Pelabelan Data Ulasan Produk *Video Games*

helpful_votes	*total_votes*	...	*review_body*	*review_date*	label
21	35	...	Yesterday i received m...	8/20/2012	tidak bermanfaat
6	7	...	Exactly as described. 5...	7/12/2014	tidak bermanfaat
...	...	...	...	...	...
21	34	...	Who pays 4 dollars m...	8/20/2012	bermanfaat
13	15	...	You get what you pay f...	8/20/2012	bermanfaat

Dari kedua hasil pelabelan tersebut, maka ulasan-ulasan yang ada pada tiap *dataset* telah memiliki *label/class*. *Label/class* tersebut memiliki nilai yaitu "bermanfaat" atau "tidak bermanfaat". Nilai tersebut nantinya akan dilakukan prediksi untuk mengetahui apakah suatu ulasan dapat dikatakan bermanfaat atau tidak bermanfaat. Adapun distribusi dari *class* pada tiap *dataset*-nya adalah sebagaimana Tabel 4.5 di bawah.

Tabel 4.5 Distribusi *Class Dataset* Hasil *Data Labeling*

Class	*Dataset* Ulasan Produk	
	Kecantikan	*Video Games*
Bermanfaat	17.254	16.264
Tidak Bermanfaat	1.031.321	1.032.311

Tiap *dataset* yang telah diberikan *label* memiliki distribusi *class* yang tidak seimbang (*class imbalance*). Seperti pada *dataset* produk kecantikan di Tabel 4.5 memiliki distribusi *class* dengan rasio 1:60. Sedangkan pada *dataset* produk *video games* memiliki rasio 1:64. Nilai rasio tersebut bermakna apabila terdapat 60 ulasan

yang ada maka terdapat satu ulasan yang memiliki label “bermanfaat”, sedangkan sisanya adalah ulasan yang memiliki label “tidak bermanfaat”.

4.3 Data Preprocessing

Data preprocessing diperlukan untuk mendapatkan data yang bersih sehingga dapat diolah untuk tahapan selanjutnya, yaitu *feature extraction*. Beberapa proses dari *data preprocessing* adalah sebagai berikut:

4.3.1 Data Cleansing

Data cleansing merupakan proses awal dalam pembersihan data ulasan. Pada proses *data cleansing* dibagi menjadi dua tahapan yaitu *cleansing* satu dan *cleansing* dua sebagaimana berikut.

1. *Cleansing* Satu

Pada tahapan *cleansing* satu adalah untuk mendapatkan data ulasan yang dapat diandalkan (valid). Data ulasan yang valid dalam penelitian ini yaitu tidak adanya ulasan yang kosong, tidak adanya ulasan yang tidak membeli produk, dan tidak adanya ulasan yang duplikat. Pada Tabel 4.6 merupakan data ulasan sebelum tahapan *cleansing* satu dengan menggunakan contoh *dataset* produk kecantikan.

Tabel 4.6 Data Ulasan Sebelum *Cleansing* Satu

customer_id	*product_id*	...	*helpful_votes*	*total_votes*	*verified_purchase*	*review_body*	...
2078375	b00nmx7kg4	...	0	1	Y	NaN	...
1115661	b00nv6lgda	...	3	4	Y	NaN	...
1797882	b001anoooe	...	0	0	Y	Love this, e...	...
18381298	b0016j22eq	...	0	0	Y	The great t...	...
50844713	b00qw14bxi	...	5	6	N	Most mask...	...
31499430	b00vok0v7m	...	10	27	N	Looking at...	...
25240884	b013j5bxfs	...	4	5	N	I love this g...	...
35751967	b00mgvka4c	...	NaN	NaN	NaN	NaN	...
22393078	b00461f4pa	...	18	23	Y	First off, I'l...	...
22234474	b00b1rpq14	...	5	5	Y	Love this th...	...
46053231	b00ocjz0ge	...	25	30	Y	This is a gr...	...
46849857	b006yc55te	...	3	5	Y	This might...	...
46849857	b006yc55te	...	3	4	Y	This might...	...

Data ulasan dari kedua *dataset* yang telah diberikan label sebelumnya masih belum didapatkannya ulasan yang valid. Seperti pada Tabel 4.6 masih terdapat ulasan yang tidak memiliki nilai pada variabel-variabelnya seperti pada variabel $review\_body$. Lalu adapun ulasan yang tidak terverifikasi telah membeli produk tersebut (*verified_purchase* bernilai "N"). Dan juga terdapat ulasan yang duplikat. Sehingga tujuan utama dari tahapan *cleansing* satu adalah untuk mengatasi hal tersebut, yang mana hasilnya adalah sebagaimana Tabel 4.7 berikut.

Tabel 4.7 Data Ulasan Setelah *Cleansing* Satu

customer_id	*product_id*	...	*helpful_votes*	*total_votes*	*verified_purchase*	*review_body*	...
1797882	b001anoooe	...	0	0	Y	Love this, e...	...
18381298	b0016j22eq	...	0	0	Y	The great t...	...
22393078	b00461f4pa	...	18	23	Y	First off, I'l...	...
22234474	b00b1rpq14	...	5	5	Y	Love this th...	...
46053231	b00ocjz0ge	...	25	30	Y	This is a gr...	...
46849857	b006yc55te	...	3	5	Y	This might...	...

Pada Tabel 4.7 data ulasan yang tidak valid telah dihilangkan. Seperti ulasan yang tidak memiliki nilai pada variabel $review\_body$ atau $helpful\_votes$, tidak terverifikasi telah membeli produk, dan duplikat. Dari hal tersebut maka telah didapatkannya data ulasan yang valid.

Pada penghapusan data ulasan tersebut, nilai dari atribut label juga ikut terhapus sehingga berpengaruh pada jumlah distribusi *class*-nya. Distribusi *class* setelah hasil *cleansing* satu adalah sebagaimana Tabel 4.8 berikut.

Tabel 4.8 Distribusi *Class Dataset* Hasil *Cleansing* Satu

Class	*Dataset* Ulasan Produk	
	Kecantikan	*Video Games*
Bermanfaat	12.739	8.315
Tidak Bermanfaat	901.996	890.783

Pada Tabel 4.8 di atas jumlah *class* bermanfaat atau tidak bermanfaat tiap *dataset* mengalami perubahan. Perubahan tersebut diakibatkan karena jumlah data ulasannya yang ikut berkurang.

2. *Cleansing* Dua

Pada tahapan *cleansing* dua adalah untuk mendapatkan teks ulasan yang bersih secara struktunya. Teks ulasan yang bersih tersebut didapatkan dari variabel $review\_body$ yang dilakukan pemrosesan sebagaimana yang telah dijelaskan pada metodologi penelitian. Beberapa proses untuk mendapatkan teks ulasan yang bersih yaitu mengubah kalimat menjadi *lowercase* (*transform to lowercase*), menghapus *url* (*remove url*), menghapus *tags html* (*remove html tags*), memperluas kata-kata singkat dan kata-kata gaul (*expand contraction and slang words*), menghapus tanda baca, simbol, dan nomor (*remove punctuation, symbol, and number*). Hasil tahapan *cleansing* dua akan diambil contoh dari dua ulasan pada masing-masing kedua *dataset*. Tabel 4.9 berikut merupakan hasil untuk proses pertama yaitu *transform to lowecase*.

Tabel 4.9 Hasil *Transform to Lowercase*

Sebelum *Transform to Lowercase*	Setelah *Transform to Lowercase*
Very good, inexpensive brush! Bought 5 for my wife & I. 4 lasted 20 months. On the last one & ready to re-order. Can't beat the price https://www.youtube .com/watch?v=nVHP49g5IPQ.	very good, inexpensive brush! bought 5 for my wife & i. 4 lasted 20 months. on the last one & ready to re-order. can't beat the price https://www.youtube .com/watch?v=nvhp49g5ipq.
AMMMAAAZZZIIIINNNGGGGG Don't think twice just buy it. It's amazing how even after the first use what a difference it makes. I love this stuff :)	ammmaaazzziiinnngggggg don't think twice just buy it. it's amazing how even after the first use what a difference it makes. i love this stuff :)
Who pays 4 dollars more for a $20 gift card? What store doesn't sell gift cards that the extra 4 dollars sounds like a good idea?	who pays 4 dollars more for a $20 gift card? what store doesn't sell gift cards that the extra 4 dollars sounds like a good idea?
Fun game, fast delivery. No problems or complaints. Nice aqnd fast delivery. Game is in excellent condition. Brand New I believe. So it is gr8	fun game, fast delivery. no problems or complaints. nice aqnd fast delivery. game is in excellent condition. brand new i believe. so it is gr8

Kata yang memiliki huruf besar pada proses ini akan diubah menjadi huruf kecil. Hal tersebut agar kata-kata yang ada memiliki bentuk yang normal atau setara. Setelah didapatkan kata-kata yang memiliki bentuk normal selanjutnya adalah menghapus *url* pada ulasan.

Suatu ulasan terkadang memiliki *url* didalamnya. *Url* tersebut umumnya mengarah ke suatu *website* yang dirujuk oleh *customer* yang memberikan ulasan

tersebut. *Url* tidak memiliki topik didalamnya sehingga kata-katanya tidak dapat dilakukan analisis dan perlu untuk dihilangkan. Hasil dari menghilangkan *url* adalah seperti pada Tabel 4.10 berikut.

Tabel 4.10 Hasil *Remove Url*

Sebelum *Remove Url*	Setelah *Remove Url*
very good, inexpensive brush! bought 5 for my wife & i. 4 lasted 20 months. on the last one & ready to re-order. can't beat the price https://www.youtube .com/watch?v=nvhp49g5ipq.	very good, inexpensive brush! bought 5 for my wife & i. 4 lasted 20 months. on the last one & ready to re-order. can't beat the price
ammmaaazzziiinnnggggg don't think twice just buy it. it's amazing how even after the first use what a difference it makes. i love this stuff :)	ammmaaazzziiinnnggggg don't think twice just buy it. it's amazing how even after the first use what a difference it makes. i love this stuff :)
who pays 4 dollars more for a $20 gift card? what store doesn't sell gift cards that the extra 4 dollars sounds like a good idea?	who pays 4 dollars more for a $20 gift card? what store doesn't sell gift cards that the extra 4 dollars sounds like a good idea?
fun game, fast delivery. no problems or complaints. nice aqnd fast delivery. game is in excellent condition. brand new i believe. so it is gr8	fun game, fast delivery. no problems or complaints. nice aqnd fast delivery. game is in excellent condition. brand new i believe. so it is gr8

Setelah *url* yang ada pada ulasan dihilangkan, maka selanjutnya adalah menghilangkan *tags html*. *Html* umumnya menggunakan tag buka/tutup untuk menunjukan perintah pada halaman *web*. Tag buka/tutup tersebut direpresentasikan dengan simbol " < " sebagai pembuka dan simbol " /> " sebagai penutup. Sehingga secara umum untuk dapat menghapus *tags html* adalah dengan cara mengidentifikasi kedua simbol tersebut yaitu dengan menggunakan *regular expression*. Hasil dari *remove html tags* adalah sebagaimana Tabel 4.11 berikut.

Tabel 4.11 Hasil *Remove Html Tags*

Sebelum *Remove Html Tags*	Setelah *Remove Html Tags*
very good, inexpensive brush! bought 5 for my wife & i. 4 lasted 20 months. on the last one & ready to re-order. can't beat the price	very good, inexpensive brush! bought 5 for my wife i. 4 lasted 20 months. on the last one ready to re-order. can't beat the price
ammmaaazzziiinnnggggg don't think twice just buy it. it's amazing how even after the first use what a difference it makes. i love this stuff :)	ammmaaazzziiinnnggggg don't think twice just buy it. it's amazing how even after the first use what a difference it makes. i love this stuff :)

who pays 4 dollars more for a $20 gift card? what store doesn't sell gift cards that the extra 4 dollars sounds like a good idea?	who pays 4 dollars more for a $20 gift card? what store doesn't sell gift cards that the extra 4 dollars sounds like a good idea?
fun game, fast delivery. no problems or complaints. nice aqnd fast delivery. game is in excellent condition. brand new i believe. so it is gr8	fun game, fast delivery. no problems or complaints. nice aqnd fast delivery. game is in excellent condition. brand new i believe. so it is gr8

Ulasan-ulasan yang memiliki *tags html* telah dihilangkan seperti tag "
" yang merupakan *line break* dan tag "&amp" untuk *ampersand*. Tag *line break* dan *ampersand* tidak dapat dilakukan analisis pada kata-katanya sehingga perlu dihilangkan.

Suatu ulasan berkemungkinan memiliki *contraction word* dan *slang word* didalamnya. *Contraction word* merupakan kata-kata singkat seperti *"you're"*, *"i'am", dan "it's".* Sedangkan *slang word* merupakan kata-kata gaul seperti *"asap", "b2b", dan "eg".* Dari hal itu maka kedua jenis kata tersebut perlu untuk diperluas. Pada perluasannya digunakan masing-masing referensi dari (Andrew Tucker, 2014) untuk *contraction word* dan (Mbaye, 2020) untuk *slang word*. Hasil perluasannya adalah sebagaimana Tabel 4.12 berikut.

Tabel 4.12 Hasil *Expand Contraction* and *Slang Words*

Sebelum *Expand Contraction and Slang Words*	Setelah *Expand Contraction and Slang Words*
very good, inexpensive brush! bought 5 for my wife i. 4 lasted 20 months. on the last one ready to re-order. can't beat the price	very good, inexpensive brush! bought 5 for my wife i. 4 lasted 20 months. on the last one ready to re-order. can not beat the price
ammmaaazzziiinnngggggg don't think twice just buy it. it's amazing how even after the first use what a difference it makes. i love this stuff :)	ammmaaazzziiinnngggggg do not think twice just buy it. it is amazing how even after the first use what a difference it makes. i love this stuff :)
who pays 4 dollars more for a $20 gift card? what store doesn't sell gift cards that the extra 4 dollars sounds like a good idea?	who pays 4 dollars more for a $20 gift card? what store does not sell gift cards that the extra 4 dollars sounds like a good idea?
fun game, fast delivery. no problems or complaints. nice aqnd fast delivery. game is in excellent condition. brand new i believe. so it is gr8	fun game, fast delivery. no problems or complaints. nice aqnd fast delivery. game is in excellent condition. brand new i believe. so it is great

Kata-kata seperti *"can't"* dan *"don't"* pada Tabel 4.12 diperluas menjadi *"can not"* dan *"do not"*. Adapun kata *slang word* seperti *gr8* juga diperluas menjadi

great. Perluasan kata tersebut dibutuhkan untuk mengetahui bentuk asli dari kata tersebut.

Selanjutnya adalah proses penghapusan tanda baca, simbol, dan nomor. Sama halnya dengan proses *remove html tags*, penghapusan tersebut akan digunakan pengenalan pola dari teks menggunakan *reguler expression* yaitu hanya diperbolehkan abjad "a" sampai dengan "z" saja dari suatu teks. Sehingga tanda baca seperti koma (,) ataupun titik (.) akan dihilangkan dari teks ulasan. Hasilnya sebagaimana Tabel 4.13 berikut.

Tabel 4.13 Hasil *Remove Punctuation Symbol and Number*

Sebelum *Remove Punctuation Symbol and Number*	Setelah *Remove Punctuation Symbol and Number*
very good, inexpensive brush! bought 5 for my wife i. 4 lasted 20 months. on the last one ready to re-order. can not beat the price	very good inexpensive brush bought for my wife i lasted months on the last one ready to re order can not beat the price
ammmaaazzziiinnnggggg do not think twice just buy it. it is amazing how even after the first use what a difference it makes. i love this stuff :)	ammmaaazzziiinnnggggg do not think twice just buy it it is amazing how even after the first use what a difference it makes i love this stuff
who pays 4 dollars more for a $20 gift card? what store does not sell gift cards that the extra 4 dollars sounds like a good idea?	who pays dollars more for a gift card what store does not sell gift cards that the extra dollars sounds like a good idea
fun game, fast delivery. no problems or complaints. nice aqnd fast delivery. game is in excellent condition. brand new i believe. so it is great	fun game fast delivery no problems or complaints nice aqnd fast delivery game is in excellent condition brand new i believe so it is great

Setelah dilakukannnya pemrosesan yaitu sampai dengan menghapus tanda baca, simbol, dan nomor pada *cleansing* dua maka telah didapatkannya teks ulasan yang bersih secara strukturnya. Pada proses selanjutnya adalah untuk mendapatkan teks ulasan yang bersih secara isinya dan akan dilanjutkan dengan tahapan *feature extraction.*

4.3.2 Tokenization

Proses selanjutnya yaitu *tokenization* akan dilakukan pemisahan teks ulasan menjadi sebuah token kata. Pemisahan tersebut dilakukan dengan cara mengidentifikasi *whitespace* dari setiap kata pada teks ulasan. Hasil dari *tokenization* adalah pada Tabel 4.14 berikut.

Tabel 4.14 Hasil *Tokenization*

Sebelum *Tokenization*	Setelah *Tokenization*
very good inexpensive brush bought for my wife i lasted months on the last one ready to re order can not beat the price	['very', 'good', 'inexpensive', 'brush', 'bought', 'for', 'my', 'wife', 'i', 'lasted', 'months', 'on', 'the', 'last', 'one', 'ready', 'to', 're', 'order', 'can', 'not', 'beat', 'the', 'price']
ammmaaazzziiinnnggggg do not think twice just buy it it is amazing how even after the first use what a difference it makes i love this stuff	['ammmaaazzziiinnnggggg', 'do', 'not', 'think', 'twice', 'just', 'buy', 'it', 'it', 'is', 'amazing', 'how', 'even', 'after', 'the', 'first', 'use', 'what', 'a', 'difference', 'it', 'makes', 'i', 'love', 'this', 'stuff']
who pays dollars more for a gift card what store does not sell gift cards that the extra dollars sounds like a good idea	['who', 'pays', 'dollars', 'more', 'for', 'a', 'gift', 'card', 'what', 'store', 'does', 'not', 'sell', 'gift', 'cards', 'that', 'the', 'extra', 'dollars', 'sounds', 'like', 'a', 'good', 'idea']
fun game fast delivery no problems or complaints nice aqnd fast delivery game is in excellent condition brand new i believe so it is great	['fun', 'game', 'fast', 'delivery', 'no', 'problems', 'or', 'complaints', 'nice', 'aqnd', 'fast', 'delivery', 'game', 'is', 'in', 'excellent', 'condition', 'brand', 'new', 'i', 'believe', 'so', 'it', 'is', 'great']

Pada Tabel 4.14 teks ulasan yang telah dilakukan *tokenization* akan berubah menjadi sebuah token kata dimana setiap teks ulasannya akan diubah menjadi list terlebih dahulu yang menampung token-token kata tersebut. List tersebut nantinya akan digunakan untuk beberapa proses selanjutnya.

4.3.3 Stopword Removal

Setelah didapatkan list dari token-token kata pada tiap ulasan, maka selanjutnya adalah mengidentifikasi *stopword* yang ada pada teks ulasan tersebut. Dalam mengidentifikasinya akan digunakan list *stopword* dari *library* NLTK dengan pengaturan bahasa inggris. Beberapa *stopword* dari *library* tersebut adalah seperti *"i", "for", "very", "a", "is", "where", "does", dan "the"*. Pada Tabel 4.15 merupakan hasil dari *stopword removal*.

Tabel 4.15 Hasil *Stopword Removal*

Sebelum *Stopword Removal*	Setelah *Stopword Removal*
['very', 'good', 'inexpensive', 'brush', 'bought', 'for', 'my', 'wife', 'i', 'lasted', 'months', 'on', 'the', 'last', 'one', 'ready', 'to', 're', 'order', 'can', 'not', 'beat', 'the', 'price']	['good', 'inexpensive', 'brush', 'bought', 'wife', 'lasted', 'months', 'last', 'one', 'ready', 'order', 'beat', 'price']

['ammmaaazzziiinnngggggg', 'do', 'not', 'think', 'twice', 'just', 'buy', 'it', 'it', 'is', 'amazing', 'how', 'even', 'after', 'the', 'first', 'use', 'what', 'a', 'difference', 'it', 'makes', 'i', 'love', 'this', 'stuff']	['ammmaaazzziiinnngggggg', 'think', 'twice', 'buy', 'amazing', 'even', 'first', 'use', 'difference', 'makes', 'love', 'stuff']
['who', 'pays', 'dollars', 'more', 'for', 'a', 'gift', 'card', 'what', 'store', 'does', 'not', 'sell', 'gift', 'cards', 'that', 'the', 'extra', 'dollars', 'sounds', 'like', 'a', 'good', 'idea']	['pays', 'dollars', 'gift', 'card', 'store', 'sell', 'gift', 'cards', 'extra', 'dollars', 'sounds', 'like', 'good', 'idea']
['fun', 'game', 'fast', 'delivery', 'no', 'problems', 'or', 'complaints', 'nice', 'aqnd', 'fast', 'delivery', 'game', 'is', 'in', 'excellent', 'condition', 'brand', 'new', 'i', 'believe', 'so', 'it', 'is', 'great']	['fun', 'game', 'fast', 'delivery', 'problems', 'complaints', 'nice', 'aqnd', 'fast', 'delivery', 'game', 'excellent', 'condition', 'brand', 'new', 'believe', 'great']

Tabel 4.15 kata umum yang ada pada bahasa inggris telah dihilangkan dari teks ulasan. Penghilangan kata yang umum tersebut dikarenakan kata tersebut memiliki nilai informasi yang rendah dari suatu teks ulasan, sehingga untuk dapat lebih fokus pada informasi yang penting maka kata umum tersebut perlu dihilangkan.

4.3.4 Lemmatization

Sebelum dilakukannya *spelling correction* dan *feature extraction*, suatu kata pada teks ulasan perlu diubah ke bentuk dasarnya terlebih dahulu. Hal tersebut bertujuan untuk setiap kata pada teks ulasan dapat diperlakukan secara sama dan sebuah model klasifikasi dapat belajar bahwa kata tersebut memiliki konteks yang sama. Hasil dari *lemmatization* adalah pada Tabel 4.16 berikut.

Tabel 4.16 Hasil *Lemmatization*

Sebelum *Lemmatization*	Setelah *Lemmatization*
['good', 'inexpensive', 'brush', 'bought', 'wife', 'lasted', 'months', 'last', 'one', 'ready', 'order', 'beat', 'price']	['good', 'inexpensive', 'brush', 'buy', 'wife', 'last', 'month', 'last', 'one', 'ready', 'order', 'beat', 'price']
['ammmaaazzziiinnngggggg', 'think', 'twice', 'buy', 'amazing', 'even', 'first', 'use', 'difference', 'makes', 'love', 'stuff']	['ammmaaazzziiinnngggggg', 'think', 'twice', 'buy', 'amaze', 'even', 'first', 'use', 'difference', 'make', 'love', 'stuff']
['pays', 'dollars', 'gift', 'card', 'store', 'sell', 'gift', 'cards', 'extra', 'dollars', 'sounds', 'like', 'good', 'idea']	['pay', 'dollar', 'gift', 'card', 'store', 'sell', 'gift', 'card', 'extra', 'dollar', 'sound', 'like', 'good', 'idea']

['fun', 'game', 'fast', 'delivery', 'problems', 'complaints', 'nice', 'aqnd', 'fast', 'delivery', 'game', 'excellent', 'condition', 'brand', 'new', 'believe', 'great']	['fun', 'game', 'fast', 'delivery', 'problem', 'complaint', 'nice', 'aqnd', 'fast', 'delivery', 'game', 'excellent', 'condition', 'brand', 'new', 'believe', 'great']

Dalam mengubah ke bentuk dasarnya, *lemmatization* bekerja dengan cara memperhatikan konteks dari suatu kata. Seperti pada kata *"'bought"* pada Tabel 4.16 berubah menjadi *"buy"*, ataupun kata *"dollars"* berubah menjadi *"dollar"*. Pengubahan kata ke bentuk dasarnya tersebut juga akan mempermudah pada tahapan berikutnya yaitu *spelling correction*.

4.3.5 Spelling Correction

Tidak semua kata pada teks ulasan memiliki ejaan yang benar, sehingga perlu dilakukannya perbaikan pada ejaan yang salah. Pada tahapan *spelling correction* akan dilakukannya cek terlebih dahulu apakah kata tersebut masuk kedalam list dari kamus kata bahasa inggris atau tidak. List kamus kata bahasa inggris yang digunakan diambil dari *library* NLTK dimana *library* tersebut menggunakan *WordNet* sebagai kamus kata bahasa inggrisnya. Hasil dari *spelling correction* adalah sebagaimana Tabel 4.17 berikut.

Tabel 4.17 Hasil *Spelling Correction*

Sebelum *Spelling Correction*	Setelah *Spelling Correction*
['good', 'inexpensive', 'brush', 'buy', 'wife', 'last', 'month', 'last', 'one', 'ready', 'order', 'beat', 'price']	['good', 'inexpensive', 'brush', 'buy', 'wife', 'last', 'month', 'last', 'one', 'ready', 'order', 'beat', 'price']
['ammmaaazzziiinnngggggg', 'think', 'twice', 'buy', 'amaze', 'even', 'first', 'use', 'difference', 'make', 'love', 'stuff']	['amazing', 'think', 'twice', 'buy', 'amaze', 'even', 'first', 'use', 'difference', 'make', 'love', 'stuff']
['pay', 'dollar', 'gift', 'card', 'store', 'sell', 'gift', 'card', 'extra', 'dollar', 'sound', 'like', 'good', 'idea']	['pay', 'dollar', 'gift', 'card', 'store', 'sell', 'gift', 'card', 'extra', 'dollar', 'sound', 'like', 'good', 'idea']
['fun', 'game', 'fast', 'delivery', 'problem', 'complaint', 'nice', 'aqnd', 'fast', 'delivery', 'game', 'excellent', 'condition', 'brand', 'new', 'believe', 'great']	['fun', 'game', 'fast', 'delivery', 'problem', 'complaint', 'nice', 'and', 'fast', 'delivery', 'game', 'excellent', 'condition', 'brand', 'new', 'believe', 'great']

Tahapan *spelling correction* dapat memperbaiki kata yang salah dalam ejaannya. Seperti kata "*aqnd*" dan "ammmaaazzziiinnngggggg" pada Tabel 4.17

memiliki ejaan yang salah. Kata "*ammmaaazzziiinnngggggg*" merupakan kata yang mengandung karakter berulang dua kali atau lebih (*elongated word*) sedangkan pada kata "*aqnd*" merupakan kata yang *typo*. Hasil ejaan kata yang benar dari masing-masing kedua kata tersebut adalah "*and*" dan "*amazing*".

Setelah dilakukannya tahapan *data preprocessing*, maka telah didapatkannya ulasan yang valid dan bersih. Menggunakan data ulasan sebelumnya maka hasil akhir dari *data preprocessing* adalah pada Tabel 4.18 berikut.

Tabel 4.18 Hasil Akhir *Data Preprocessing*

...	*review_body*	*review_date*	label	*clean_review*
...	Very good, inexpensive brush! Bought 5 for my wife & I. 4 lasted 20 months. On the last one & ready to re-order. Can't beat the price https://www.youtube .com/watch?v=nVHP49g5IPQ.	8/31/2015	tidak bermanfaat	good inexpensive brush buy wife last month last one ready order beat price
...	AMMMAAAZZZIIINNNGGGGG Don't think twice just buy it. It's amazing how even after the first use what a difference it makes. I love this stuff :)	8/31/2015	tidak bermanfaat	amazing think twice buy amaze even first use difference make love stuff
...	Who pays 4 dollars more for a $20 gift card? What store doesn't sell gift cards that the extra 4 dollars sounds like a good idea?	8/31/2015	bermanfaat	pay dollar gift card store sell gift card extra dollar sound like good idea
...	Fun game, fast delivery. No problems or complaints. Nice aqnd fast delivery. Game is in excellent condition. Brand New I believe. So it is gr8	8/31/2015	tidak bermanfaat	fun game fast delivery problem complaint nice and fast delivery game excellent condition brand new believe great

Hasil akhir dari *data preprocessing* akan disimpan ke dalam variabel *clean_review*. Untuk selanjutnya variabel tersebut akan digunakan pada tahapan *feature extraction*.

4.4 Feature Extraction

Untuk dapat melakukan klasifikasi dengan menggunakan model *machine learning* pada teks ulasan, maka diperlukan sebuah pengenalan pola. Pengenalan

pola tersebut dapat dilakukan dengan cara mengidentifikasi *feature* yang ada pada teks ulasan. Dalam mengidentifikasinya, *feature* tersebut perlu dikeluarkan. Proses mengeluarkan *feature* tersebut dinamakan *feature extraction*. Pada penelitian ini tahapan dari *feature extraction* dibagi menjadi tiga proses yaitu ekstraksi fitur *semantic*, *structural*, dan kombinasi sebagaimana penjelasan di bawah berikut.

4.4.1 Semantic Feature

Ekstraksi fitur *semantic* merupakan proses mengeluarkan fitur yang berguna untuk mengetahui makna kata pada suatu teks. Salah satu metodenya adalah TF-IDF. Pada TF-IDF suatu kata akan diberikan bobot, dimana bobot terebut bernilai antara nol sampai dengan satu (nilai normalisasi). Suatu kata akan memiliki bobot yang berat (mendekati nilai satu) apabila kata tersebut muncul di banyak tempat pada satu dokumen teks, tetapi kemunculannya sedikit pada dokumen teks yang lain.

Untuk mengetahui hasil ekstraksi fitur *semantic* menggunakan TF-IDF maka diberikan contoh empat data ulasan yang telah dilakukan *preprocessing* sebelumnya. Contoh empat data ulasan tersebut adalah sebagaimana Tabel 4.19 berikut.

Tabel 4.19 Contoh Empat Data Ulasan

No.	*review_body*	...	*clean_review*
1	Love this, excellent sun block!!	...	love excellent sun block
2	The great thing about this cream is that it doesn't smell weird like all those chemical laden ones. I get a nice healthy un-fake looking tan that isn't orange and it makes my skin soft too.	...	great thing cream smell weird like chemical laden one get nice healthy un fake look tan orange make skin soft
3	Great Product, I'm 65 years old and this is all it claims to be!	...	great product year old claim
4	I use them as shower caps & conditioning caps. I like that they're in bulk. It saves a lot of money.	...	use shower cap condition cap like bulk save lot money

Pada dasarnya hasil dari metode TF-IDF adalah berupa *sparse matrix*. *Sparse matrix* merupakan *matrix* (*array* dua dimensi) yang didalamnya terdiri dari sebagian besar nilai nol. Sebagaimana Tabel 4.20 di bawah diperlihatkan hasil dari

TF-IDF menggunakan empat data ulasan sebelumnya (variabel $clean\_review$) yaitu apabila *sparse matrix* tersebut diubah kedalam bentuk tabel.

Tabel 4.20 Hasil Pembobotan TF-IDF

	block	bulk	cap	chemical	claim	...	weird	year
0	0.5	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000
1	0.0	0.000000	0.000000	0.227962	0.000000	...	0.227962	0.000000
2	0.0	0.000000	0.000000	0.000000	0.465162	...	0.000000	0.465162
3	0.0	0.293337	0.586674	0.000000	0.000000	...	0.000000	0.000000

Pada Tabel 4.20 merupakan hasil TF-IDF yang telah diubah bentuknya menjadi sebuah tabel. Kolom paling kiri merupakan indeks dari dokumen teks, sedangkan baris paling atas yaitu setiap kolomnya merupakan kata-kata yang telah diberikan nilai pembobotan oleh TF-IDF. Semakin banyak kata yang diberikan pembobotan oleh TF-IDF maka semakin banyak juga jumlah kolom yang dihasilkan.

Namun dalam melakukan *fitting* pada model klasifikasi, hasil dari TF-IDF yang berupa tabel tersebut tidak dapat digunakan. Untuk dapat menggunakannya maka hasil TF-IDF tersebut perlu diubah terlebih dahulu ke bentuk awalnya yaitu berupa *sparse matrix* yang dikompresi. *Sparse matrix* yang dikompresi tersebut menghilangkan nilai nol yang ada pada TF-IDF sehingga fokus pada nilai hasil pembobotan sebagaimana pada Tabel 4.21 berikut.

Tabel 4.21 Hasil TF-IDF Untuk *Fitting* Model

Indeks Dokumen dan kata	Nilai Pembobotan TF-IDF
(2, 10)	0.3667390112974172
(3, 18)	0.29333722228100645
(3, 15)	0.29333722228100645
(3, 24)	0.29333722228100645
(3, 1)	0.29333722228100645
(3, 5)	0.29333722228100645
(3, 2)	0.5866744445620129
(3, 25)	0.29333722228100645
(3, 33)	0.29333722228100645
(3, 13)	0.23127043535458108

Pada Tabel 4.21 tersebut diambil dari 10 kata terakhir yang diberikan pembobotan TF-IDF. Terdapat tiga variabel nilai yang ada yaitu pada bagian paling kiri merupakan indeks dari dokumen teks yang dirujuk. Sedangkan bagian tengah merupakan indeks dari token kata yang dilakukan perhitungan TF-IDF. Terakhir pada bagian kanan merupakan nilai hasil pembobotan TF-IDF. Nilai hasil pembobotan tersebutlah yang akan digunakan dalam proses *training* maupun *testing* pada model klasifikasi nantinya.

4.4.2 Structural Feature

Pada ekstraksi fitur *structural* merupakan proses dalam mengeluarkan fitur yang berfungsi untuk mengetahui struktur dan format dari sebuah dokumen teks. Salah satu struktur dan format dari sebuah dokumen teks tersebut adalah jumlah kata dan jumlah kalimat. Dari hal tersebut maka untuk dapat melakukan ekstraksi fitur *structural* perlu dilakukan perhitungan pada jumlah kata dan jumlah kalimat yang ada pada tiap ulasan. Dengan menggunakan contoh empat data ulasan sebelumnya maka hasil ekstraksi fitur *structural* adalah sebagaimana Tabel 4.22 berikut.

Tabel 4.22 Hasil Ektraksi Fitur *Structural*

	review_body	...	jumlah_kata	jumlah_kalimat
0	Love this, excellent sun block!!	...	5	2
1	The great thing about this cream is that it doesn't smell weird like all those chemical laden ones. I get a nice healthy un-fake looking tan that isn't orange and it makes my skin soft too.	...	39	2
2	Great Product, I'm 65 years old and this is all it claims to be!	...	14	1
3	I use them as shower caps & conditioning caps. I like that they're in bulk. It saves a lot of money.	...	21	3

Perhitungan jumlah kata dan jumlah kalimat pada Tabel 4.22 tersebut adalah berdasarkan variabel *review_body*. Sebagaimana pada metodologi penelitian, pada perhitungan jumlah katanya ulasan tersebut dilakukan *preprocessing* terlebih dahulu yaitu menggunakan tahapan *cleansing* dua untuk menghilangkan simbol dan tanda baca yang ada pada teks ulasan. Sedangkan pada perhitungan jumlah

kalimatnya adalah dengan menggunakan fungsi *sent_tokenize* pada *library* NLTK. Fungsi tersebut bekerja dengan cara mengidentifikasi tanda baca titik yang ada pada teks ulasan. Dari hasil perhitungan tersebut maka kedua kolom yaitu jumlah kata dan jumlah kalimat dapat digunakan sebagai data *training* dan *testing* pada model klasifikasi nantinya.

4.4.3 Combined Feature

Berdasarkan saran dari penelitian sebelumnya, maka pada penelitian ini akan digabungkan fitur-fitur yang sebelumnya telah diusulkan yaitu fitur *semantic* dan fitur *structural*. Hal tersebut dilakukan dikarenakan kedua fitur tersebut memiliki perbedaan pandangan dalam memahami teks ulasan dimana pada fitur *semantic* adalah memedulikan makna kata yang ada pada teks ulasan, sedangkan pada fitur *structural* adalah tidak memedulikan makna kata tersebut. Dengan masih menggunakan contoh empat data ulasan sebelumnya maka hasil penggabungan kedua fitur tersebut adalah seperti pada Tabel 4.23 berikut.

Tabel 4.23 Hasil Penggabungan Kedua *Feature*

	block	*bulk*	*cap*	...	*weird*	*year*	jumlah _kata	jumlah _kalimat
0	0.5	0.000000	0.000000	...	0.000000	0.000000	5.0	2.0
1	0.0	0.000000	0.000000	...	0.227962	0.000000	39.0	2.0
2	0.0	0.000000	0.000000	...	0.000000	0.465162	14.0	1.0
3	0.0	0.293337	0.586674	...	0.000000	0.000000	21.0	3.0

Apabila diubah kedalam bentuk tabel maka Tabel 4.23 tersebut merupakan hasil penggabungan dari kedua fitur yang telah diusulkan. Pada hasil penggabungannya, fitur *semantic* terletak pada bagian kiri tabel sedangkan fitur *structural* terletak pada bagian kanan tabel. Kolom paling kiri merupakan indeks dari dokumen teks, sedangkan pada tiap selnya merupakan nilai dari hasil pembobotan TF-IDF serta perhitungan jumlah kata dan jumlah kalimat pada tiap ulasannya.

Dari hasil penggabungan tersebut, untuk dapat melakukan *fitting* pada model maka hasil ekstraksi fitur *structural* perlu diubah terlebih dahulu menyesuaikan hasil dari ekstraksi fitur *semantic*. Adapun setelah dilakukannya

pengubahan tersebut nilai fitur *semantic* dan fitur *structural* akan disamakan, yaitu dengan cara melakukan *scaling* pada data. Pada *scaling* datanya akan digunakan *scaling* dari *StandardScaler*. Sehingga hasil dari pengubahan, penggabungan, serta penyamaan nilai untuk kedua fitur tersebut adalah seperti pada Tabel 4.24 berikut.

Tabel 4.24 Hasil Penggabungan Kedua *Feature* Untuk *Fitting* Model

Indeks Dokumen dan kata/kolom	Nilai TF-IDF serta Jumlah Kata dan Kalimat Hasil Scaling Data
(3, 1)	2.3094010767585034
(3, 2)	2.3094010767585034
(3, 5)	2.3094010767585034
(3, 13)	2.216232705278345
(3, 15)	2.3094010767585034
(3, 18)	2.3094010767585034
(3, 24)	2.3094010767585034
(3, 25)	2.3094010767585034
(3, 33)	2.3094010767585034
(3, 36)	1.6830321893718083
(3, 37)	4.242640687119285

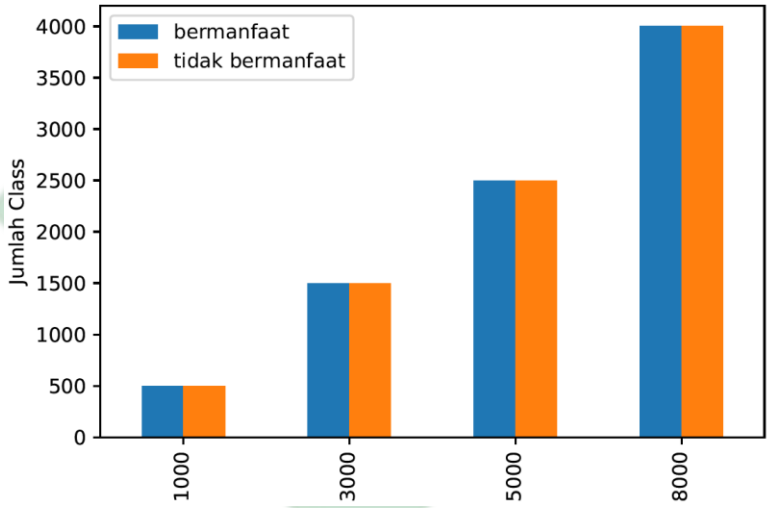
Pada *StandardScaler* data akan dilakukan *scaling* dengan cara menghitung *mean* dan *variance* pada data. Namun dikarenakan data hasil penggabungan pada fitur kombinasi adalah berupa *sparse matrix*, maka pada *scaling* datanya tidak dapat digunakan perhitungan *mean* dikarenakan dapat mengubah bentuk asli datanya. Sehingga nilai-nilai pada Tabel 4.24 tersebut merupakan nilai-nilai yang dapat digunakan sebagai data *training* dan *testing* pada model klasifikasi nantinya.

4.5 Modeling

Setelah tahapan *feature extraction*, selanjutnya adalah tahapan *modeling*. Pada tahapan *modeling*, kedua fitur yang telah diekstrak sebelumnya akan digunakan untuk membangun model klasifikasi. Model klasifikasi yang digunakan adalah model SVM dengan pengaturan *default kernel* yaitu "rbf". Model tersebut akan digunakan untuk melakukan prediksi pada data ulasan yang baru untuk menentukan apakah ulasan tersebut masuk kedalam ulasan yang bermanfaat atau tidak bermanfaat. Beberapa proses dari tahapan *modeling* adalah sebagaimana penjelasan di bawah berikut.

4.5.1 Pembagian Data

Data yang digunakan untuk tahapan *modeling* akan dibagi terlebih dahulu. Pembagian data tersebut juga merupakan skenario untuk mengevaluasi performa model menggunakan kedua fitur yang telah diusulkan. Serta dikarenakan jumlah data yang ada terlalu besar sehingga dapat menyebabkan proses *modeling* menjadi lama. Pada pembagian datanya akan digunakan data dari masing-masing kedua *dataset* sejumlah 1000, 3000, 5000, dan 8000 baris ulasan. Model klasifikasi juga sensitif terhadap *high imbalance class*. Sehingga pada pembagian datanya juga akan disetarakan jumlah dari *class*-nya. Ilustrasi dari pembagian data tersebut adalah sebagaimana Gambar 4.1 berikut.



Gambar 4.1 Hasil Pembagian Data

Pada masing-masing pembagian datanya, tiap data akan memiliki jumlah *class* yang seimbang. Seperti pada pembagian data untuk 1000 baris ulasan, jumlah *class* yang dimilikinya didapatkan dari masing-masing 500 baris pertama untuk *class* bermanfaat dan 500 baris pertama untuk *class* tidak bermanfaat. Hal tersebut juga berlaku pada pembagian data lainnya yaitu 3000, 5000, dan 8000 baris.

4.5.2 Model Evaluation

Pada proses *model evaluation*, model akan diuji performanya menggunakan metode *10-fold cross validation* yang selanjutnya dilakukan penilaian berdasarkan nilai *accuracy*, *precision*, *recall*, dan *f-measure*. Adapun *confusion matrix* digunakan untuk mengetahui apakah model dapat membedakan *True Positive* dan *True Negative* dengan baik. Untuk skenarionya adalah berdasarkan pada kedua fitur

yang telah diusulkan sebelumnya dan hasil dari pembagian data. Berikut merupakan hasil evaluasinya.

1. Semantic Feature

Pada skenario pertama akan digunakan ekstraksi fitur *semantic* untuk menguji model. Hasil dari pengujian model tersebut adalah sebagaimana Tabel 4.25 berikut.

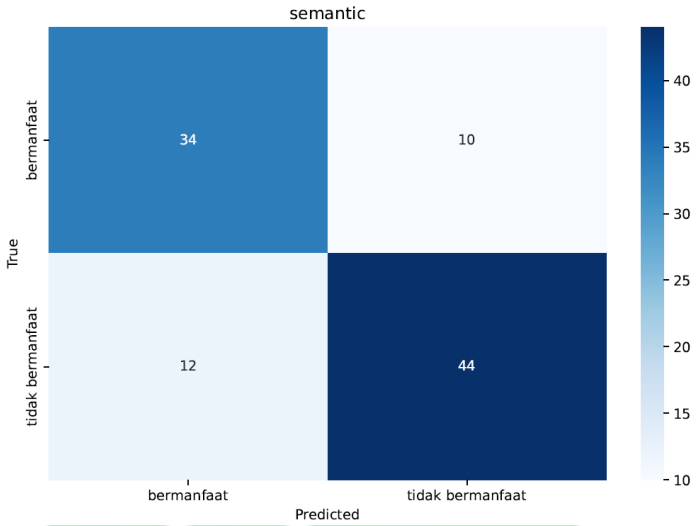
Tabel 4.25 Hasil Pengujian *Semantic Feature*

Jumlah Data dan Kata	Rata-rata Hasil *10-fold Cross Validation*			
	Accuracy	*Precision*	*Recall*	*F-measure*
Dataset Ulasan Produk Kecantikan				
1000 data × 4433 kata	0.728 ± 0.04	0.735 ± 0.04	0.731 ± 0.04	0.725 ± 0.04
3000 data × 7090 kata	0.762 ± 0.02	0.762 ± 0.02	0.762 ± 0.02	0.761 ± 0.02
5000 data × 8881 kata	0.768 ± 0.02	0.769 ± 0.02	0.769 ± 0.02	0.768 ± 0.02
8000 data × 10.902 kata	0.775 ± 0.01	0.775 ± 0.01	0.775 ± 0.01	0.774 ± 0.01
Dataset Ulasan Produk *Video Games*				
1000 data × 5741 kata	0.798 ± 0.03	0.808 ± 0.03	0.800 ± 0.03	0.794 ± 0.03
3000 data × 9824 kata	0.809 ± 0.02	0.811 ± 0.02	0.809 ± 0.02	0.809 ± 0.02
5000 data × 12.565 kata	0.821 ± 0.01	0.822 ± 0.01	0.822 ± 0.01	0.821 ± 0.01
8000 data × 15.741kata	0.826 ± 0.01	0.826 ± 0.01	0.826 ± 0.01	0.825 ± 0.01

Pada Tabel 4.25 di atas merupakan nilai rata-rata dari hasil iterasi masing-masing skor penilaian pada *10-fold cross validation*. Pada *dataset* ulasan produk kecantikan masing-masing skor penilaian tersebut mengalami kenaikan seiring dengan banyaknya data yang digunakan. Adapun nilai *f-measure* tertinggi pada *dataset* tersebut adalah sebesar 0.774.

Lalu pada *dataset* ulasan produk *video games* juga mengalami peningkatan seiring dengan banyaknya data yang digunakan. Adapun pada *dataset* tersebut memiliki nilai skor masing-masing lebih tinggi jika dibandingkan dengan *dataset* ulasan produk kecantikan. Untuk nilai *f-measure* tertinggi-nya adalah sebesar 0.825.

Pada *confusion matrix* akan digunakan hasil satu iterasi saja dari *10-fold cross validation*. Hasilnya adalah sebagaimana Gambar 4.2 berikut.



Gambar 4.2 *Confusion Matrix Semantic Feature*

Seperti pada Gambar 4.2 dengan menggunakan fitur *semantic* model tidak hanya dapat memprediksi dengan benar kelas positif saja (label bermanfaat) namun juga dapat dengan benar memprediksi kelas negatif (label tidak bermanfat). Hal tersebut terbukti pada nilai TP dan TN masing-masing 34 dan 44.

2. Structural Feature

Pada skenario kedua akan digunakan ekstraksi fitur *structural* berupa jumlah kata dan jumlah kalimat untuk menguji performa dari model. Hasil pengujiannya adalah sebagaimana Tabel 4.26 berikut.

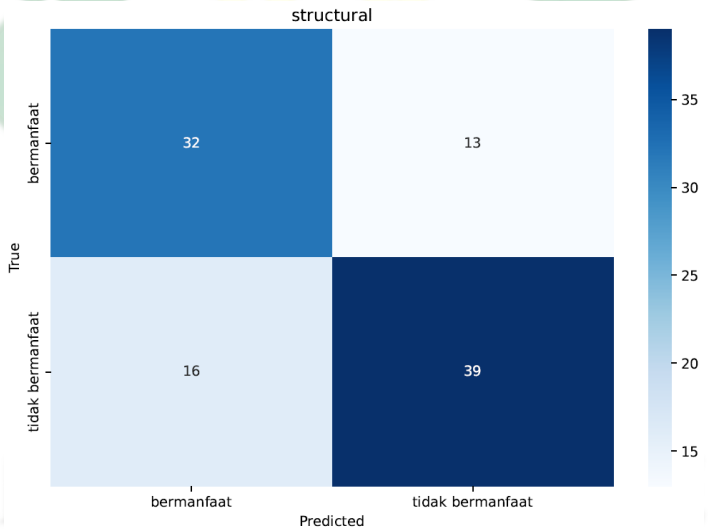
Tabel 4.26 Hasil Pengujian *Structural Feature*

Jumlah Data	Rata-rata Hasil *10-fold Cross Validation*			
	Accuracy	*Precision*	*Recall*	*F-measure*
Dataset Ulasan Produk Kecantikan				
1000 data	0.777 ± 0.03	0.778 ± 0.03	0.777 ± 0.03	0.776 ± 0.03
3000 data	0.777 ± 0.01	0.777 ± 0.01	0.777 ± 0.01	0.776 ± 0.01
5000 data	0.779 ± 0.02	0.779 ± 0.02	0.779 ± 0.02	0.779 ± 0.02
8000 data	0.780 ± 0.01	0.780 ± 0.01	0.780 ± 0.01	0.780 ± 0.01
Dataset Ulasan Produk *Video Games*				
1000 data	0.807 ± 0.05	0.811 ± 0.05	0.805 ± 0.05	0.803 ± 0.05
3000 data	0.817 ± 0.01	0.819 ± 0.01	0.817 ± 0.01	0.816 ± 0.01
5000 data	0.821 ± 0.01	0.823 ± 0.01	0.821 ± 0.01	0.820 ± 0.01

8000 data	0.823 ± 0.01	0.825 ± 0.01	0.823 ± 0.01	0.823 ± 0.01

Pada tiap skenarionya, hasil pengujian fitur *structural* menghasilkan nilai *f-measure* relatif lebih tinggi jika dibandingkan dengan fitur *semantic*. Seperti pada *dataset* ulasan produk kecantikan nilai *f-measure* tertinggi-nya adalah sebesar 0.780. Namun untuk nilai *f-measure* tertinggi dari *dataset* ulasan produk *video games* pada hasil pengujian fitur *structural* masih dibawah hasil pengujian fitur *semantic* yaitu bernilai 0.823.

Lalu sama halnya dengan hasil pengujian fitur *semantic*, pada dataset ulasan produk *video games* untuk hasil pengujian fitur *structural*-nya juga memiliki nilai skor masing-masing lebih tinggi jika dibandingkan dengan *dataset* ulasan produk kecantikan. Adapun hasil *confusion matrix* untuk satu iterasi pada *10-fold cross validation*-nya adalah sebagaimana Gambar 4.3 berikut.



Gambar 4.3 *Confusion Matrix Structural Feature*

Pada Gambar 4.3 hasil *confusion matrix* untuk fitur *structural* TP dan TN memiliki nilai masing-masing 32 dan 39. Hal tersebut membuktikan bahwa dengan menggunakan fitur *structural* model dapat melakukan klasifikasi tidak hanya memprediksi dengan benar kelas positif namun juga dapat memprediksi dengan benar kelas negatif.

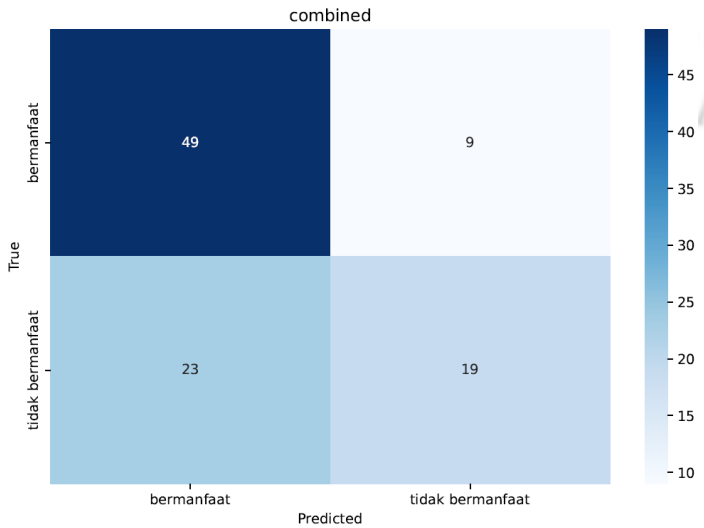
3. Combined Feature

Skenario terakhir yaitu akan digunakan fitur kombinasi yang merupakan penggabungan antara fitur *semantic* dan fitur *structural*. Hasil dari pengujian fitur kombinasi adalah pada Tabel 4.27 berikut.

Tabel 4.27 Hasil Pengujian *Combined Feature*

Jumlah Data	Rata-rata Hasil *10-fold Cross Validation*			
	Accuracy	*Precision*	*Recall*	*F-measure*
Dataset Ulasan Produk Kecantikan				
1000 data	0.696 ± 0.05	0.712 ± 0.04	0.696 ± 0.05	0.686 ± 0.05
3000 data	0.713 ± 0.03	0.717 ± 0.02	0.713 ± 0.02	0.711 ± 0.02
5000 data	0.720 ± 0.01	0.722 ± 0.01	0.719 ± 0.01	0.718 ± 0.01
8000 data	0.737 ± 0.01	0.740 ± 0.01	0.737 ± 0.01	0.736 ± 0.01
Dataset Ulasan Produk *Video Games*				
1000 data	0.731 ± 0.05	0.755 ± 0.06	0.731 ± 0.06	0.723 ± 0.06
3000 data	0.758 ± 0.02	0.766 ± 0.02	0.759 ± 0.02	0.756 ± 0.02
5000 data	0.772 ± 0.01	0.776 ± 0.01	0.771 ± 0.01	0.771 ± 0.01
8000 data	0.786 ± 0.01	0.789 ± 0.01	0.786 ± 0.01	0.785 ± 0.01

Hasil pengujian fitur kombinasi memiliki nilai rata-rata dibawah fitur *semantic* dan fitur *structural*. Dimana nilai *f-measure* tertinggi untuk masing-masing *dataset* adalah sebesar 0.736 dan 0.785. Adapun hasil *confusion matrix* dari fitur kombinasi tersebut adalah seperti pada Gambar 4.4 berikut.



Gambar 4.4 *Confusion Matrix Combined Feature*

Hasil *confusion matrix* tersebut tidak jauh berbeda dengan hasil *confusion matrix* pada fitur *semantic* atau fitur *structural*. Namun berdasarkan *confusion matrix*-nya model dengan menggunakan fitur kombinasi lebih cenderung dapat memprediksi dengan benar kelas positif, meskipun juga dapat dengan benar memprediksi kelas negatif-nya. Hal tersebut dibuktikan dengan nilai TP dan TN masing-masing adalah 49 dan 19.

4.6 Analisis Hasil

Bagian ini membahas terkait hasil dari penelitian yang sedang diteliti saat ini serta membandingkannya dengan hasil penelitian yang telah dilakukan sebelumnya sehingga dapat ditemukan pengetahuan baru yang dapat menjadi referensi untuk penelitian selanjutnya.

Pada penelitian ini di semua skenario yang telah dilakukan sebelumnya, *dataset* ulasan produk *video games* (*search product*) memiliki nilai *f-measure* yang lebih tinggi jika dibandingkan dengan *dataset* ulasan produk kecantikan (*experience product*). Hal tersebut sesuai dengan penelitian Muhammad Shahid Iqbal Malik (2020) yang mengatakan bahwa indikator isi ulasan lebih berpengaruh pada *search product* dalam memprediksi ulasan yang bermanfaat jika dibandingan dengan *experience product*. Penelitian Chua & Banerjee (2016) juga mengatakan bahwa *search product* pada dasarnya lebih mudah untuk dilakukan penilaian pada kualitas produknya bahkan sebelum produk tersebut dibeli dibandingkan dengan *experience product*.

Pada penelitian Du et al., (2019) dengan menggunakan *dataset*, model, dan *features* yang sama, penggunaan fitur *structural*-nya menghasilkan nilai akurasi yang lebih rendah jika dibandingkan dengan fitur *semantic*. Hal tersebut tidak sesuai dengan penelitian ini dimana di banyak skenario yang telah dilakukan, fitur *structural* menghasilkan nilai akurasi yang relatif lebih tinggi jika dibandingkan dengan fitur *semantic*. Adapun perbandingan antara penelitian ini dengan penelitian Du et al., tersebut sebagaimana Tabel 4.28 berikut.

Tabel 4.28 Perbandingan Penelitian

Nama Peneliti	*Dataset*	Model	*Features*	Nilai Akurasi
Du et al.,		SVM	*Semantic* (*unigram* TF-IDF)	0.749

	Amazon ulasan produk *video games*		*Structural* (*count word*)	0.647
			Structural (*count sentence*)	0.639
Penelitian ini			*Semantic* (TF-IDF)	0.825
			Structural (*count word + count sentence*)	0.823

Pada Tabel 4.28 terdapat perbedaan nilai akurasi dimana pada penelitian Du et al., dengan menggunakan 23.100 baris ulasan menghasilkan nilai akurasi sebesar 0.749 untuk fitur *semantic* sedangkan pada penelitian ini dengan fitur yang sama menghasilkan nilai akurasi sebesar 0.825. Adapun dengan menggabungkan perhitungan jumlah kata dan jumlah kalimat pada fitur structural yaitu pada penelitian ini menghasilkan nilai akurasi yang lebih baik jika dibandingkan dengan penelitian Du et al., dengan nilai akurasi yang didapatkan adalah sebesar 0.823.

Penelitian ini juga menggabungkan antara fitur *semantic* dan fitur *structural* berdasarkan saran dari penelitian sebelumnya. Pada hasil penggabungan tersebut didapatkan nilai rata-rata yang lebih rendah jika dibandingkan dengan fitur *semantic* dan fitur *structural* meskipun telah dilakukannnya *scaling* pada data.

Selanjutnya kinerja model SVM pada penelitian ini juga terbukti menghasilkan performa klasifikasi yang tinggi. Hal tersebut dibuktikan dengan nilai dari *f-measure* yang didapatkan serta hasil *confusion matrix*-nya.

BAB V
PENUTUP

5.1 Kesimpulan

Berdasarkan dengan hasil yang telah didapatkan dan dijelaskan sebelumnya. Maka pada penelitian ini dapat diambil kesimpulan sebagaimana berikut:

1. Ekstraksi *semantic feature* dapat dilakukan dengan cara menggunakan metode TF-IDF. Dimana teks ulasan dilakukan *preprocessing* terlebih dahulu untuk menghilangkan *stopword*, simbol, nomor, dan tanda baca sehingga didapatkan teks ulasan yang bersih. Teks ulasan yang bersih tersebut selanjutnya dilakukan pembobotan oleh TF-IDF untuk mendapatkan nilai yang dapat digunakan pada tahapan *modeling*.
2. Ekstraksi *structural feature* dapat dilakukan dengan cara melakukan perhitungan jumlah kata dan jumlah kalimat dari teks ulasan. Pada perhitungan jumlah kata, teks ulasan dilakukan *preprocessing* terlebih dahulu untuk menghilangkan karakter atau simbol yang bukan termasuk kata. Sedangkan untuk perhitungan jumlah kalimat dilakukan dengan cara mengidentifikasi tanda baca titik pada ulasan sebagai akhir dari suatu kalimat.
3. Performa model SVM untuk klasifikasi ulasan bermanfaat memiliki nilai *f-measure* paling tinggi bernilai 0.825 untuk fitur *semantic*. Sedangkan pada fitur *structural* nilai *f-measure* paling tinggi bernilai 0.823 lalu pada fitur kombinasi nilai *f-measure* paling tinggi bernilai 0.785. Model juga terbukti dapat memprediksi dengan benar kelas positif dan kelas negatif-nya berdasarkan pada hasil *confusion matrix*.

5.2 Saran

Dari penelitian mengenai Klasifikasi Kualitas Ulasan Produk Berdasarkan *Semantic* dan *Structural Features* Menggunakan *Support Vector Machine* ini tentu masih jauh dari kata sempurna, sehingga perlu dilakukannya pengembangan lebih lanjut. Adapun rekomendasi serta saran perbaikan untuk penelitian selanjutnya berdasarkan hasil temuan yang didapatkan, yaitu:

1. Menambahkan validasi hasil pelabelan yang dilakukan oleh pakar secara manual dengan cara mengambil sampel dari keseluruhan data yang telah dilabelkan sebelumnya. Dikarenakan pada penelitian ini pelabelan masih dilakukan secara otomatis yaitu dengan berdasarkan pada penelitian terdahulu.
2. Menggunakan *scaling* data lainnya yaitu selain *StandarScaler*. Dikarenakan *StandarScaler* tidak dapat digunakan untuk menghitung *mean* dari data yang berupa *sparse matrix* dikompresi.

DAFTAR PUSTAKA

Altınel, B., & Ganiz, M. C. (2018). Semantic text classification: A survey of past and recent advances. *Information Processing and Management*, *54*(6), 1129–1153. https://doi.org/10.1016/j.ipm.2018.08.001

Alzu'Bi, S., Alsmadiv, A., Alqatawneh, S., Al-Ayyoub, M., Hawashin, B., & Jararweh, Y. (2019). A Brief Analysis of Amazon Online Reviews. *2019 6th International Conference on Social Networks Analysis, Management and Security, SNAMS 2019*, 555–560. https://doi.org/10.1109/SNAMS.2019.8931816

Andrew Tucker. (2014). *Common Contractions*. *San José State University Writing Center*. Retrieved from https://www.sjsu.edu/writingcenter/docs/handouts/Contractions.pdf

Bahassine, S., Madani, A., Al-Sarem, M., & Kissi, M. (2020). Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University - Computer and Information Sciences*, *32*(2), 225–231. https://doi.org/10.1016/j.jksuci.2018.05.010

Balakumar, J., & Mohan, S. V. (2019). Artificial bee colony algorithm for feature selection and improved support vector machine for text classification. *Information Discovery and Delivery*, *47*(3), 154–170. https://doi.org/10.1108/IDD-09-2018-0045

Cao, Q., Duan, W., & Gan, Q. (2011). Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems*, *50*(2), 511–521. https://doi.org/10.1016/j.dss.2010.11.009

Chakraborty, U., & Bhat, S. (2018). The Effects of Credible Online Reviews on Brand Equity Dimensions and Its Consequence on Consumer Behavior. *Journal of Promotion Management*, *24*(1), 57–82. https://doi.org/10.1080/10496491.2017.1346541

Chandra, M. A., & Bedi, S. S. (2021). Survey on SVM and their application in image classification. *International Journal of Information Technology (Singapore)*, *13*(5). https://doi.org/10.1007/s41870-017-0080-1

Cheng, C. H., & Chen, H. H. (2019). Sentimental text mining based on an additional features method for text classification. *PLoS ONE*, *14*(6), 1–17. https://doi.org/10.1371/journal.pone.0217591

Chou, S. Y., Picazo-Vela, S., & Pearson, J. M. (2013). The Effect of Online Review Configurations, Prices, and Personality on Online Purchase Decisions: A Study of Online Review Profiles on eBay. *Journal of Internet Commerce*, *12*(2), 131–153. https://doi.org/10.1080/15332861.2013.817862

Chua, A. Y. K., & Banerjee, S. (2016a). Helpfulness of user-generated reviews as a function of review sentiment, product type and information quality. *Computers in Human Behavior*, *54*, 547–554. https://doi.org/10.1016/j.chb.2015.08.057

Chua, A. Y. K., & Banerjee, S. (2016b). Helpfulness of user-generated reviews as a function of review sentiment, product type and information quality. *Computers in Human Behavior*, *54*, 547–554. https://doi.org/10.1016/j.chb.2015.08.057

Dang, S., & Ahmad, P. H. (2014). Text Mining : Techniques and its Application Text Mining View project Text Mining: Techniques and its Application. *IJETI International Journal of Engineering & Technology Innovations*, *1*(March). Retrieved from www.ijeti.com

Dash, A., Zhang, D., & Zhou, L. (2021). Personalized Ranking of Online Reviews Based on Consumer Preferences in Product Features. *International Journal of Electronic Commerce*, *25*(1), 29–50. https://doi.org/10.1080/10864415.2021.1846852

Donthu, N., Kumar, S., Pandey, N., Pandey, N., & Mishra, A. (2021). Mapping the electronic word-of-mouth (eWOM) research: A systematic review and bibliometric analysis. *Journal of Business Research*, *135*(February), 758–773. https://doi.org/10.1016/j.jbusres.2021.07.015

Du, J., Rong, J., Michalska, S., Wang, H., & Zhang, Y. (2019). Feature selection for helpfulness prediction of online product reviews: An empirical study. *PLoS ONE*, *14*(12), 1–26. https://doi.org/10.1371/journal.pone.0226902

Herr, P. M., Kardes, F. R., & Kim, J. (1991). Effects of Word-of-Mouth and Product-Attribute Information on Persuasion: An Accessibility-Diagnosticity Perspective. *Journal of Consumer Research*, *17*(4), 454. https://doi.org/10.1086/208570

Hládek, D., Staš, J., & Pleva, M. (2020). Survey of automatic spelling correction. *Electronics (Switzerland)*, *9*(10), 1–29. https://doi.org/10.3390/electronics9101670

Hsu, B. M. (2020). Comparison of supervised classification models on textual data. *Mathematics*, *8*(5). https://doi.org/10.3390/MATH8050851

Huang, A. H., Chen, K., Yen, D. C., & Tran, T. P. (2015). A study of factors that contribute to online review helpfulness. *Computers in Human Behavior*, *48*, 17–27. https://doi.org/10.1016/j.chb.2015.01.010

Huang, Z., & Benyoucef, M. (2013). From e-commerce to social commerce: A close look at design features. *Electronic Commerce Research and Applications*, *12*(4), 246–259. https://doi.org/10.1016/j.elerap.2012.12.003

Isabelle, G., Jason, W., Stephen, B., & Vladimir, V. (2002). Gene Selection for

Cancer Classification using Support Vector Machines. *Machine Learning*, *46*, 389–422. Retrieved from http://link.springer.com/article/10.1023/A:1012487302797

Kadhim, A. I. (2018). An Evaluation of Preprocessing Techniques for Text Classification. *International Journal of Computer Science and Information Security*, *16*(6), 22–32. Retrieved from https://sites.google.com/site/ijcsis/

Khorsheed, M. S., & Al-Thubaity, A. O. (2013). Comparative evaluation of text classification techniques using a large diverse Arabic dataset. *Language Resources and Evaluation*, *47*(2), 513–538. https://doi.org/10.1007/s10579-013-9221-8

Kim, H. joon, Kim, J., Kim, J., & Lim, P. (2018). Towards perfect text classification with Wikipedia-based semantic Naïve Bayes learning. *Neurocomputing*, *315*, 128–134. https://doi.org/10.1016/j.neucom.2018.07.002

Kim, S. J., Maslowska, E., & Malthouse, E. C. (2018). Understanding the effects of different review features on purchase probability. *International Journal of Advertising*, *37*(1), 29–53. https://doi.org/10.1080/02650487.2017.1340928

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). *Text Mining in Organizational Research*. *Organizational Research Methods* (Vol. 21). https://doi.org/10.1177/1094428117722619

Korfiatis, N., García-Bariocanal, E., & Sánchez-Alonso, S. (2012). Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications*, *11*(3), 205–217. https://doi.org/10.1016/j.elerap.2011.10.003

Krishnamoorthy, S. (2015). Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, *42*(7), 3751–3759. https://doi.org/10.1016/j.eswa.2014.12.044

Lackermair, G., Kailer, D., & Kanmaz, K. (2013). Importance of Online Product Reviews from a Consumer's Perspective. *Advances in Economics and Business*, *1*(1), 1–5. https://doi.org/10.13189/aeb.2013.010101

Lee, S., & Choeh, J. Y. (2014). Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Systems with Applications*, *41*(6), 3041–3046. https://doi.org/10.1016/j.eswa.2013.10.034

Lourdusamy, R., & Abraham, S. (2018). A Survey on Text Pre-processing Techniques and Tools. *International Journal of Computer Sciences and Engineering*, *06*(03), 148–157. https://doi.org/10.26438/ijcse/v6si3.148157

Mak, G., & Montreal, G. M. M. (2000). *the Implementation of Support Vector Machines Using the Sequential Minimal Optimization Algorithm*.

Malik, M. S.I., & Hussain, A. (2017). Helpfulness of product reviews as a function of discrete positive and negative emotions. *Computers in Human Behavior*, *73*, 290–302. https://doi.org/10.1016/j.chb.2017.03.053

Malik, M. S.I., & Hussain, A. (2018). An analysis of review content and reviewer variables that contribute to review helpfulness. *Information Processing and Management*, *54*(1), 88–104. https://doi.org/10.1016/j.ipm.2017.09.004

Malik, Muhammad Shahid Iqbal. (2020). Predicting users' review helpfulness: the role of significant review and reviewer characteristics. *Soft Computing*, *24*(18), 13913–13928. https://doi.org/10.1007/s00500-020-04767-1

Markoulidakis, I., Kopsiaftis, G., Rallis, I., & Georgoulas, I. (2021). Multi-Class Confusion Matrix Reduction method and its application on Net Promoter Score classification problem. *ACM International Conference Proceeding Series*, 412–419. https://doi.org/10.1145/3453892.3461323

Mbaye, M. (2020). Up-to-date list of Slangs for Text Preprocessing. Retrieved July 29, 2022, from https://www.kaggle.com/code/nmaguette/up-to-date-list-of-slangs-for-text-preprocessing/notebook

Meng, Y., Yang, N., Qian, Z., & Zhang, G. (2021). What makes an online review more helpful: An interpretation framework using xgboost and shap values. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(3), 466–490. https://doi.org/10.3390/jtaer16030029

Mirończuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, *106*, 36–54. https://doi.org/10.1016/j.eswa.2018.03.058

Moreno-Llamas, A., García-Mayor, J., & De la Cruz-Sánchez, E. (2020). The impact of digital technology development on sitting time across Europe. *Technology in Society*, *63*(May). https://doi.org/10.1016/j.techsoc.2020.101406

Moreno-Torres, J. G., Saez, J. A., & Herrera, F. (2012). Study on the impact of partition-induced dataset shift on k-fold cross-validation. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(8), 1304–1312. https://doi.org/10.1109/TNNLS.2012.2199516

Moussa, M., & Măndoiu, I. I. (2018). Single cell RNA-seq data clustering using TF-IDF based methods. *BMC Genomics*, *19*(Suppl 6). https://doi.org/10.1186/s12864-018-4922-4

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on amazon.com. *MIS Quarterly: Management Information Systems*, *34*(1), 185–200. https://doi.org/10.2307/20721420

Ngo-Ye, T. L., & Sinha, A. P. (2014). The influence of reviewer engagement characteristics on online review helpfulness: A text regression model.

Decision Support Systems, *61*(1), 47–58. https://doi.org/10.1016/j.dss.2014.01.011

Pai, M. Y., Chu, H. C., Wang, S. C., & Chen, Y. M. (2013). Electronic word of mouth analysis for service experience. *Expert Systems with Applications*, *40*(6), 1993–2006. https://doi.org/10.1016/j.eswa.2012.10.024

Pan, Y., & Zhang, J. Q. (2011). Born Unequal: A Study of the Helpfulness of User-Generated Product Reviews. *Journal of Retailing*, *87*(4), 598–612. https://doi.org/10.1016/j.jretai.2011.05.002

Refaeilzadeh, P., Tang, L., Liu, H., Angeles, L., & Scientist, C. D. (2020). Encyclopedia of Database Systems. *Encyclopedia of Database Systems*. https://doi.org/10.1007/978-1-4899-7993-3

Rizwan, A., Iqbal, N., Ahmad, R., & Kim, D. H. (2021). Wr-svm model based on the margin radius approach for solving the minimum enclosing ball problem in support vector machine classification. *Applied Sciences (Switzerland)*, *11*(10). https://doi.org/10.3390/app11104657

Salehi, F., Abdollahbeigi, B., Langroudi, A. C., & Salehi, F. (2012). The Impact of Website Information Convenience on E-commerce Success of Companies. *Procedia - Social and Behavioral Sciences*, *57*, 381–387. https://doi.org/10.1016/j.sbspro.2012.09.1201

Schuckert, M., Liu, X., & Law, R. (2016). Stars, Votes, and Badges: How Online Badges Affect Hotel Reviewers. *Journal of Travel and Tourism Marketing*, *33*(4), 440–452. https://doi.org/10.1080/10548408.2015.1064056

Wan, C. H., Lee, L. H., Rajkumar, R., & Isa, D. (2012). A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine. *Expert Systems with Applications*, *39*(15), 11880–11888. https://doi.org/10.1016/j.eswa.2012.02.068

Wei, L., Wei, B., & Wang, B. (2012). Text Classification Using Support Vector Machine with Mixture of Kernel. *Journal of Software Engineering and Applications*, *05*(12), 55–58. https://doi.org/10.4236/jsea.2012.512b012

Weisstein, F. L., Song, L., Andersen, P., & Zhu, Y. (2017). Examining impacts of negative reviews and purchase goals on consumer purchase decision. *Journal of Retailing and Consumer Services*, *39*(August), 201–207. https://doi.org/10.1016/j.jretconser.2017.08.015

Wu, J. (2017). Review popularity and review helpfulness: A model for user review effectiveness. *Decision Support Systems*, *97*, 92–103. https://doi.org/10.1016/j.dss.2017.03.008

Zamir, A., Khan, H. U., Mehmood, W., Iqbal, T., & Akram, A. U. (2020). A feature-centric spam email detection model using diverse supervised machine

learning algorithms. *Electronic Library*, *38*(3), 633–657. https://doi.org/10.1108/EL-07-2019-0181

Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications*, *41*(4 PART 1), 1476–1482. https://doi.org/10.1016/j.eswa.2013.08.044

zhou, Y., Yang, S., li, yixiao, chen, Y., Yao, J., & Qazi, A. (2020). Does the review deserve more helpfulness when its title resembles the content? Locating helpful reviews by text mining. *Information Processing and Management*, *57*(2), 102179. https://doi.org/10.1016/j.ipm.2019.102179