



INSTITUT TEKNOLOGI INDONESIA

ANALISIS SENTIMEN PADA VIDEO ULASAN PRODUK  
KECANTIKAN MENGGUNAKAN *RANDOM FOREST CLASSIFIER*

SKRIPSI

ANNISSA MUTIAPUTRI

1151600004

INFORMATIKA  
TANGERANG SELATAN

2021



INSTITUT TEKNOLOGI INDONESIA

ANALISIS SENTIMEN PADA VIDEO ULASAN PRODUK  
KECANTIKAN MENGGUNAKAN *RANDOM FOREST CLASSIFIER*

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar  
Sarjana Komputer

ANNISSA MUTIAPUTRI

1151600004

INFORMATIKA  
TANGERANG SELATAN

2021

**HALAMAN PERNYATAAN ORISINALITAS**

Skripsi ini adalah hasil karya saya sendiri,  
dan semua sumber baik yang dikutip maupun dirujuk  
Telah saya nyatakan dengan benar.

Nama : Annissa Mutiaputri

NPM : 1151600004

Tanda Tangan :

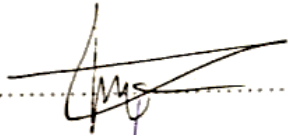
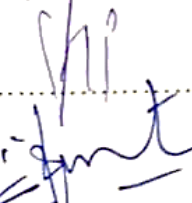
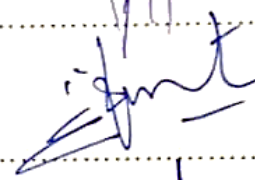

Tanggal : 17 Februari 2021

## LEMBAR PENGESAHAN

Skripsi ini diajukan oleh :  
Nama : Annissa Mutiaputri  
NPM : 1151600004  
Program Studi : Informatika  
Judul Skripsi : Analisis Sentimen Pada Video Ulasan Produk Kecantikan Menggunakan *Random Forest Classifier*

Telah berhasil dipertahankan di hadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Komputer pada Program Studi Informatika Institut Teknologi Indonesia


## DEWAN PENGUJI

Pembimbing : Dino Hariatma P, M.Kom (.....)  
Penguji 1 : Dra. Sulistyowati, M.Kom (.....)  
Penguji 2 : Dra. Endang RD, M.Kom (.....)  
Penguji 3 : Husni, M.Kom, MSc (.....)

Ditetapkan di : Kampus Institut Teknologi Indonesia, Tangerang Selatan

Tanggal : 17 Februari 2021

## KETUA PROGRAM STUDI INFORMATIKA

  
(Dra. Sulistyowati, M.Kom)

## KATA PENGANTAR

Segala puji dan syukur kami panjatkan kehadiran Allah SWT yang telah melimpahkan rahmat dan hidayah-Nya kepada kita semua. *Shalawat* serta salam semoga tercurah kepada Rasulullah SAW beserta keluarganya-Nya karena berkat rohmat, hidayah, dan ridho-Nya sehingga tugas akhir ini dapat diselesaikan dengan baik sebagai syarat untuk menyelesaikan Program Sarjana (S1) pada Fakultas Teknologi Industri, Jurusan Informatika di Institut Teknologi Indonesia.

Dalam penelitian Tugas Akhir ini, bantuan, bimbingan, saran serta semangat dari berbagai pihak sangat membantu dalam menyelesaikan tugas akhir ini. Oleh karena itu terimakasih yang tak terhingga diucapkan kepada :

1. Pak Dino Hariatma, M.Kom, sebagai Dosen Pembimbing yang telah menyediakan waktu, tenaga, dan pikiran untuk mengarahkan penyusunan Tugas Akhir ini;
2. Ibu Dra. Sulistyowati, M.Kom, sebagai Ketua Program Studi Informatika yang telah mengarahkan dalam penyusunan Tugas Akhir ini;
3. Ibu Melani Indriasari, M.Kom, sebagai Dosen Pembimbing Akademik yang telah mengarahkan dari awal semester sampai saat ini;
4. Kedua orangtua, Ibu Rina Fatimah dan Pak Bramantyo Sukarno atas doa, motivasi, ridho, dan segala hal yang telah diberikan dalam mendukung perkuliahan;
5. Teman-teman 999+, Indira Farhana Pramesti, Jasmine Kharila Salma, Nadia Asy-Syaffa, Nyoman Dinda Suhartini, Salma Hananinayah, dan Titania Herliawati yang selalu memberikan semangat serta motivasi selama penyusunan tugas akhir ini;
6. Teman-teman seperjuangan skripsi, Azmi Refani, Nove Suarti Hia, Veronica Yudistiana, Fairuz Zahirah, Candra, Nur Afifa Halimatul Sadiyah, Nur Indah Wulandari, Sari Lavenia Tampubolon, dan M. Yogi Mustofa yang telah memberikan semangat serta dukungan dalam menyelesaikan tugas akhir ini;
7. Teman-teman Informatika ITI angkatan 2016 dan Bengkel Seni ITI yang telah membantu, mendukung, dan memotivasi dalam hari-hari perkuliahan;
8. Pemilik tempat fotokopi Indah *Copy Centre* Setu yang telah mendukung dalam pembuatan tugas akhir ini;

Akhir kata, saya berharap Tuhan Yang Maha Esa berkenan membalas kebaikan semua pihak yang telah membantu. Semoga skripsi ini membawa manfaat bagi pengembangan ilmu.

**HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI  
TUGAS AKHIR / SKRIPSI UNTUK KEPENTINGAN AKADEMIS**

Sebagai sivitas akademika Institut Teknologi Indonesia, saya yang bertanda tangan di bawah ini :

Nama : Annissa Mutiাপুତ્રි

NPM : 1151600004

Program Studi : Informatika

Jenis Karya : Tugas Akhir/Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Institut Teknologi Indonesia Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalty Free Right*) atas karya ilmiah saya yang berjudul :

Analisis Sentimen Pada Video Ulasan Produk Kecantikan Menggunakan *Random Forest Classifier*

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Institut Teknologi Indonesia berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan Tugas Akhir/Skripsi saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

## ABSTRAK

**Nama** : Annissa Mutiapusri  
**Program Studi** : Informatika  
**Judul** : Analisis Sentimen Pada Video Ulasan Produk Kecantikan Menggunakan *Random Forest Classifier*  
**Dosen Pembimbing** : Dino Hariatma, M.Kom.

Banyak sekali video ulasan produk kecantikan yang tersebar di *platform* YouTube, namun sangat sedikit yang menganalisis sentiment pada video ulasan tersebut. Penelitian ini melakukan analisis terhadap video ulasan produk kecantikan menggunakan *Random Forest Classifier*. Data dikumpulkan dari *platform* YouTube. Data yang masuk kemudian diseleksi agar tidak terjadi duplikasi data dan berisi hanya ulasan dari satu jenis produk kecantikan yang tidak lebih dari 12 menit. Video dari data yang telah diseleksi akan diunduh dan diubah menjadi bentuk teks melalui dua tahap, yaitu tahap pengubahan video menjadi audio, dan tahap pengubahan audio menjadi teks. Video yang telah menjadi teks kemudian diseleksi kembali agar data yang gagal direkognasi dapat dihapus. Labelisasi dilakukan secara manual dengan label *rating* dari angka 1 hingga angka 5. Data teks yang telah dilabelisasi selanjutnya akan dilakukan *Preprocessing* agar data dapat dimasukkan ke dalam model. Pelatihan data menggunakan *Random Forest Classifier* dengan banyak *tree* sejumlah 5. Hasil akhir analisis adalah klasifikasi *rating* data uji dengan nilai akurasi model sebesar 0.5 atau 50%.

**Kata Kunci** : Klasifikasi *Rating*, *Random Forest Classifier*, Video Ulasan Produk Kecantikan, YouTube.

## ABSTRACT

*There are so many review videos of beauty product in YouTube platform, but there are only few people that analyze the sentiment of video reviews. This research is analyzing the beauty product's review videos using Random Forest Classifier. Data collected from YouTube platform. The data that has been collected will be selected to prevent data duplication and to keep only one type of product's reviews that are no longer than 12 minutes. The selected data's video will be downloaded and converted into text by two steps, the first one is to convert the video into audio, and the second one is to convert the audio into text. The converted video will be selected to delete the data that have recognitions failure. The labeling is done manually by adding the 'rate' label from 1 to 5. The labeled data will be preprocessed so it can be processed by the model. The data was trained using Random Forest Classifier with 5 trees. The result of analysis is the rate classification of test data with accuracy point 0.5 or 50%.*

**Keywords** : Rate Classification, Random Forest Classifier, Video Review of Beauty Products, YouTube.

## DAFTAR ISI

HALAMAN JUDUL.....	i
HALAMAN PERNYATAAN ORISINALITAS .....	ii
LEMBAR PENGESAHAN.....	ii
KATA PENGANTAR .....	iv
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI .....	v
ABSTRAK .....	vi
DAFTAR GAMBAR .....	ix
DAFTAR TABEL.....	x
BAB 1 PENDAHULUAN.....	1
1.1. Latar Belakang.....	1
1.2. Rumusan Masalah.....	3
1.3. Tujuan dan Manfaat Penelitian.....	3
1.4. Ruang Lingkup Penelitian.....	3
1.5. Metodologi Penelitian .....	3
1.6. Sistematika Penelitian .....	4
BAB 2 LANDASAN TEORI .....	6
2.1. Natural Language Processing .....	6
2.1.1. Analisis Sentimen .....	6
2.1.2. Text Preprocessing.....	7
2.1.3. Bag of Words.....	8
2.2. Machine Learning .....	8
2.3. Random Forest Classifier .....	10
2.3.1. Decision <i>Tree</i> .....	10
2.3.2. Algoritma Random Forest.....	12
2.4. Confusion matrix .....	14
2.4.1. Precission .....	14
2.4.2. Recall .....	14
2.4.3. Akurasi .....	14
BAB 3 ANALISIS DAN PERANCANGAN.....	15
3.1. Analisis Model.....	15
3.1.1. Analisis Masalah.....	15
3.1.2. Analisis Kebutuhan.....	15
3.2. Perancangan Model Analisis Sentimen.....	16
3.2.1. Pengumpulan Video.....	16
3.2.2. Pengubahan Video Menjadi Teks.....	18



3.2.3. Pengklasifikasian Data.....	20
BAB 4 IMPLEMENTASI DAN PENGUJIAN.....	28
4.1. Implementasi .....	28
4.1.1. Lingkungan Eksperimen .....	28
4.1.2. Dataset.....	29
4.1.3. Implementasi Model Analisis Sentimen .....	30
4.2. Pengujian.....	34
4.3. Analisis.....	36
4.3.1. Analisis Studi Kasus .....	36
4.3.2. Analisis Model.....	36
BAB 5 PENUTUP .....	37
5.1. Kesimpulan.....	37
5.2. Saran .....	37

## DAFTAR GAMBAR

Gambar 2.1 Diagram Algoritma Umum dan <i>Machine Learning</i> .....	9
Gambar 2.2 <i>Decision Tree</i> .....	10
Gambar 2.3 <i>Random Forest</i> .....	12
Gambar 3.1. Bagan Solusi Permasalahan.....	15
Gambar 3.2. Bagan Proses Pengumpulan Video .....	16
Gambar 3.3. Bagan Proses Perubahan Video Menjadi Teks.....	18
Gambar 3.4. Bagan Proses Pengubahan Video Menjadi Audio .....	19
Gambar 3.5 Bagan Proses Mengubah Audio Menjadi Teks .....	19
Gambar 3.6. Bagan Tahap <i>Preprocessing</i> .....	21
Gambar 3.7. Perbandingan Data Uji dan Data Latih .....	24
Gambar 3.8 Pembentukan <i>root tree</i> .....	26
Gambar 3.9 pembentukan <i>tree sample</i> .....	26
Gambar 3.10 Alur <i>Random Forest</i> .....	27
Gambar 4.1 Implementasi <i>Preprocessing – Raw Data</i> .....	30
Gambar 4.2 Implementasi <i>Preprocessing – Preprocessed Data</i> .....	30
Gambar 4.3 Pembentukan matriks <i>bag of words</i> .....	30
Gambar 4.4 Pembagian Data Latih dan Data Uji .....	31
Gambar 4.5 Hasil Klasifikasi Model .....	32

## DAFTAR TABEL

Tabel 2.1 <i>Confusion Matrix</i> .....	14
Tabel 3.1. Hasil <i>Scraping</i> Data.....	17
Tabel 3.2. Hasil Penyeleksian Judul Video .....	18
Tabel 3.3 Contoh Hasil Rekognasi Audio menjadi Teks .....	19
Tabel 3.4. Contoh Hasil Labelisasi .....	20
Tabel 3.5. Contoh Normalisasi Kata.....	22
Tabel 3.6. Contoh Stopwords .....	22
Tabel 3.7 Contoh Stem Kata .....	22
Tabel 3.8 Contoh Matriks <i>Bag of Words</i> .....	23
Tabel 3.9 Sampel Data Membuat <i>Tree</i> .....	25
Tabel 3.10 Perhitungan Gini Index dan Gini <i>Split</i> .....	25
Tabel 4.1 Distribusi <i>Class</i> Dataset.....	29
Tabel 4.2 Percobaan Akurasi Jumlah <i>tree</i> pertama .....	32
Tabel 4.3 Percobaan Akurasi Jumlah <i>tree</i> kedua.....	32
Tabel 4.4 <i>Confusion Matrix</i> Pengujian .....	32
Tabel 4.5 Pengujian Model .....	35

## **BAB 1**

### **PENDAHULUAN**

#### **1.1. Latar Belakang**

YouTube merupakan salah satu *platform* yang biasa digunakan untuk memberikan kebebasan dalam menyampaikan pendapat dan membagikan momen ke seluruh dunia dalam bentuk video dan audio. Pada tahun 2020 belakangan ini YouTube mengalami kenaikan jumlah penonton akibat pandemi virus *Covid-19*. Kebijakan *lockdown* di beberapa daerah di Indonesia membuat tidak sedikit orang beralih profesi menjadi *content creator* di YouTube. Selain meningkatnya *content creator* di *platform* YouTube, jumlah penonton pada *platform* YouTube juga meningkat.

*Content creator* atau yang biasa disebut sebagai *Vlogger* memiliki nama masing-masing untuk setiap kategori. *Beauty Vlogger* merupakan salah satu konten kreator yang membuat konten di bidang kecantikan. *Beauty Vlogger* biasa mengunggah konten seperti tips perawatan wajah, cara *makeup*, serta ulasan mengenai produk-produk kecantikan sedetil mungkin. Video ulasan produk kecantikan pun semakin banyak beredar di *platform* YouTube. Selain semakin banyaknya *beauty vlogger*, jumlah pencarian kata kunci pada kategori *Beauty and Fitness* di YouTube juga meningkat sebanyak 25% dari lima tahun terakhir (Google Trends, 2021).

*Vlog* sebagai bukti perkembangan media digital, turut membuka cara baru untuk menghasilkan *PR coverage* (aktifitas untuk mempromosikan barang pada media *online* maupun *offline*) dan sangat mampu diandalkan untuk memberikan pengaruh (*influence*) serta menarik perhatian dari sejumlah konsumen (Mariezka, Haufar, & Yustikasari, 2018). Keberadaan *Beauty Vlogger* dapat dimanfaatkan sehingga dapat memberikan manfaat dan keuntungan bagi para produsen produk kecantikan. Tanpa perlu memberikan imbalan, produk kecantikan akan diulas secara rinci oleh para *Beauty Vlogger*. Para *beauty vlogger* biasanya akan mengulas produk kecantikan mulai dari harga, kemasan, kekentalan krim, kesan pertama yang dirasakan ketika menggunakan produk, hingga hasil pemakaian produk setelah beberapa hari. Hal ini tentu saja akan menjadi pertimbangan bagi orang-orang yang menonton ulasan tersebut apakah akan membeli dan menggunakan produk kecantikan tersebut atau tidak. Besarnya pengaruh *Beauty Vlogger* pada penjualan produk kecantikan membuat para produsen produk kecantikan patut mengetahui sentimen dari isi video produk kecantikan.

*Machine Learning* dapat digunakan untuk mengetahui sentimen dari video ulasan produk kecantikan. *Machine Learning* (ML) atau pembelajaran mesin merupakan pendekatan dalam kecerdasan buatan yang banyak digunakan untuk menggantikan atau menirukan perilaku manusia untuk menyelesaikan masalah atau melakukan otomatisasi (Bhandary, 2019). *Machine Learning* biasa digunakan untuk melakukan klasifikasi dan prediksi. Pelatihan diperlukan oleh *Machine Learning* untuk memprediksi dan mengklasifikasikan data dengan baik. Klasifikasi adalah metode dalam *Machine Learning* yang digunakan oleh mesin untuk memilah atau mengklasifikasikan objek berdasarkan ciri tertentu sebagaimana manusia mencoba membedakan benda satu dengan yang lainnya (Bhandary, 2019). Prediksi digunakan untuk menerka keluaran dari suatu data masukan berdasarkan data yang sudah dipelajari dalam *training* (Bhandary, 2019). Dengan bantuan *Machine Learning*, data ulasan produk kecantikan dapat diberikan klasifikasi sentimennya secara otomatis. Terdapat enam algoritma *Machine Learning* yang biasa digunakan untuk menganalisis sentimen, algoritma tersebut adalah algoritma *Logistic Regression*, *K-nearest Neighbors*, *Support Vector Machine* (SVM), *Naïve Bayes*, *Decision Tree*, dan *Random Forest Classification*.

*Random Forest* adalah kombinasi dari masing-masing *tree* yang baik kemudian dikombinasikan ke dalam suatu model (Ahmad, Maqsood, Khan, & Jehad, 2012). *Random Forest Classification* terbentuk dari banyak *Decision Tree* dengan tujuan untuk mengklasifikasikan data. Model akan dilatih dengan cara membentuk beberapa *Decision Tree*. Masing-masing *Decision Tree* akan menghasilkan satu hasil klasifikasi data. Seluruh hasil dari setiap *tree* kemudian akan dihitung dan diambil hasil yang paling banyak. Sistem *voting* tersebut membuat algoritma *Random Forest Classification* memiliki nilai akurasi yang tinggi.

Berdasarkan penelitian analisis sentimen pada isi video yang telah dilakukan sebelumnya oleh Unnathi Bandhary pada tahun 2019, Bandhary mmeneliti tentang pendeteksian ujaran kebencian pada video menggunakan *Machine Learning*. Pada penelitiannya, Bandhary melakukan klasifikasi video ujaran kebencian menggunakan empat buah algoritma, yaitu algoritma Multinomial Naïve Bayes, Linear SVM, Random Forest, dan RNN. Berdasarkan penelitiannya tersebut, *Random Forest Classifier* menghasilkan hasil klasifikasi pada isi video dengan akurasi yang jauh lebih baik dibandingkan dengan algoritma lain.

## 1.2. Rumusan Masalah

Berdasarkan latar belakang masalah di atas, masalah pokok yang akan dibahas adalah sebagai berikut :

- Apakah *Random Forest Classifier* dapat digunakan untuk mengklasifikasikan video ulasan produk kecantikan ke dalam 5 *class*.
- Seberapa baik performa algoritma *Random Forest Classification* pada analisis sentimen video ulasan produk kecantikan.

## 1.3. Tujuan dan Manfaat Penelitian

Tujuan dan manfaat dari penelitian ini adalah sebagai berikut :

- Mengetahui performa algoritma *Random Forest Classification* dalam mengklasifikasikan video ulasan produk kecantikan ke dalam 5 *class rating*.
- Membantu produsen produk kecantikan dalam mendapatkan *rating* produk dari video ulasan produk kecantikan.

## 1.4. Ruang Lingkup Penelitian

Agar pengerjaan tugas akhir ini menjadi lebih terarah dan mendapatkan hasil yang lebih baik, maka penelitian akan dibatasi pada ruang lingkup sebagai berikut :

- Video yang digunakan berdurasi minimal satu menit dan maksimal dua belas menit lima puluh sembilan detik.
- Video yang digunakan adalah video dengan bahasa Indonesia.
- Produk kecantikan yang akan dibahas terbatas pada produk perawatan wajah seperti pembersih wajah, *toner*, *serum*, dan krim wajah.
- Video yang digunakan hanya memuat ulasan satu merek dagang.
- Video yang digunakan hanya memuat satu jenis produk perawatan wajah.
- Bahasa pemrograman yang digunakan adalah *Python*.
- Data akan diklasifikasikan menjadi lima *class rating*, yaitu *class 1* dengan ulasan positif (*rating 5*), *class 2* dengan ulasan agak positif (*rating 4*), *class 3* dengan ulasan netral (*rating 3*), *class 4* dengan ulasan agak negatif (*rating 2*), *class 5* dengan ulasan negatif (*rating 1*).

## 1.5. Metodologi Penelitian

Metodologi penelitian yang digunakan dalam tugas akhir ini adalah sebagai berikut:

a) Studi Pustaka

Pada tahap ini dilakukan pengumpulan referensi yang diperlukan dalam penelitian. Hal ini dilakukan untuk mengumpulkan informasi yang dibutuhkan dalam penulisan penelitian ini. Referensi dapat berupa buku, jurnal, artikel, atau situs internet yang berhubungan dengan analisis sentimen video ulasan produk kecantikan menggunakan *Random Forest Classifier*.

b) Analisis dan Perancangan

Pada tahap ini hal-hal yang dibutuhkan untuk penelitian akan dianalisis berdasarkan ruang lingkup penelitian, kemudian dirancang sehingga menghasilkan model yang baik.

c) Implementasi

Pada tahap ini, pengumpulan dan pembentukan dataset serta implementasi model dilakukan berdasarkan hasil analisis dan perancangan menggunakan bahasa pemrograman *Python*.

d) Pengujian

Pada tahap ini model yang telah dirancang dan diimplementasikan akan diuji coba apakah model dapat digunakan untuk mengklasifikasikan video ulasan produk kecantikan menggunakan *Random Forest Classifier* dengan baik.

e) Dokumentasi

Pada tahap ini penelitian yang telah dilakukan akan didokumentasikan mulai dari tahap analisis dan perancangan sampai tahap pengujian dalam bentuk laporan tugas akhir.

## 1.6. Sistematika Penelitian

Secara garis besar penulisan laporan tugas akhir ini terbagi menjadi beberapa bagian. Bagian-bagian laporan tugas akhir ini disusun sehingga menghasilkan susunan laporan seperti berikut :

- BAB 1 : berisi Pendahuluan penelitian yang terdiri dari latar belakang, rumusan masalah, tujuan dan manfaat penelitian, ruang lingkup penelitian, metodologi penelitian yang digunakan, serta sistematika penulisan penelitian.
- BAB 2 : berisi Landasan Teori penelitian yang terdiri dari teori dasar *Natural Language Processing*, *Machine Learning*, *Random Forest Classifier* dan *Confusion Matrix*.
- BAB 3 : berisi Analisis dan Perancangan yang meliputi analisis model, analisis masalah, analisis kebutuhan, dan perancangan model analisis sentimen.

- BAB 4 : berisi Implementasi dan Pengujian yang terdiri dari pengimplementasian lingkungan eksperimen, dataset, model analisis sentimen, pengujian model, serta analisis yang terbagi menjadi dua yaitu analisis studi kasus dan analisis model.
- BAB 5 : mencantumkan tentang Penutup yang berisi simpulan dari pembahasan analisis sentimen pada video ulasan produk kecantikan menggunakan *Random Forest* dan saran untuk pengembangannya.
- Daftar referensi mencantumkan beberapa acuan referensi/pustaka yang digunakan untuk membuat laporan ini.



## **BAB 2**

### **LANDASAN TEORI**

#### **2.1. Natural Language Processing**

*Natural Language Processing* adalah sebuah set metode yang bertujuan untuk membuat bahasa manusia dapat dipahami oleh komputer (Einstein, 2019). *Natural Language Processing* adalah cabang ilmu komputer dan linguistik yang mengkaji interaksi antara komputer dan bahasa alami manusia (Ruben, 2020). Beberapa prinsip dasar pada *natural language processing*, yaitu (Ruben, 2020):

- **Fonetik/Fonologi**

Fonologi secara harfiah berarti ilmu bunyi. Secara definisi, fonologi adalah cabang linguistik atau ilmu bahasa yang mengkaji bunyi ujar dalam bahasa tertentu.

- **Morfologi**

Morfologi adalah cabang linguistik yang mengidentifikasi satuan-satuan dasar bahasa sebagai satuan gramatikal. Morfologi mempelajari seluk-beluk bentuk kata serta pengaruh perubahan-perubahan bentuk kata terhadap golongan dan arti kata.

- **Sintaksis**

Sintaksis adalah bagian dari ilmu bahasa yang mempelajari prinsip dan peraturan dalam membuat kalimat.

- **Semantik**

Semantik berarti memberikan tanda. Semantik adalah cabang dari ilmu bahasa yang mempelajari makna yang terkandung dalam suatu bahasa.

- **Pragmatik**

Pragmatik adalah ilmu bahasa yang mempelajari hubungan antara konteks dan makna. Pragmatik mengkaji kondisi-kondisi penggunaan bahasa manusia yang ditentukan oleh konteks kemasyarakatan.

##### **2.1.1. Analisis Sentimen**

Sentimen analisis (*opinion mining*) adalah metode *natural language processing* yang digunakan untuk menentukan apakah suatu data memiliki sentimen yang bersifat positif, negatif, atau netral. Sentimen analisis biasa digunakan dalam data teks untuk membantu para pebisnis memonitor sentimen dari produk milik suatu *brand* dalam ulasan pelanggan dan mengetahui apa yang pelanggan butuhkan (MonkeyLearn).

Sentimen analisis memiliki beberapa tipe, diantaranya adalah sebagai berikut (MonkeyLearn) :

- *Fine-grained Sentiment Analysis*

*Fine grained sentiment analysis* menganalisis sentiment berdasarkan polaritasnya. Kategori polaritas dapat disesuaikan dengan kebutuhan bisnis. Contoh kategori polaritas yang dapat digunakan adalah sangat positif, positif, netral, *negatif*, dan sangat negatif. Fine-grained sentiment analysis juga dapat digunakan dalam pemberian *5-star ratings* dalam sebuah ulasan.

- Deteksi Emosi

Sentimen analisis tipe ini bertujuan untuk mendeteksi emosi seperti bahagia, frustrasi, marah, sedih, dan sebagainya. Sebagian besar sistem deteksi emosi menggunakan leksikon (list kata dan emosi) atau algoritma *Machine Learning* yang kompleks.

### 2.1.2. Text Preprocessing

*Preprocessing* merupakan tahapan awal dalam mengolah data input sebelum memasuki proses tahapan utama dari metode *Latent Semantic Analysis* (LSA). *Preprocessing text* dilakukan untuk tujuan penyeragaman dan kemudahan pembacaan serta proses LSA selanjutnya (Aji P., Baizal SSi. and Firdaus S.T., 2011). *Preprocessing* terdiri dari beberapa tahapan. Adapun tahapan *preprocessing* (Triawati, 2009) , yaitu:

- *Case Folding*

*Case folding* merupakan tahapan yang mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf ‘a’ sampai dengan ‘z’ yang diterima. Karakter selain huruf dihilangkan dan dianggap *delimiter* (pembatas)(Triawati, 2009).

- *Tokenizing*

Tahap *tokenizing* atau *parsing* adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya (Triawati, 2009)

- *Filtering*

Tahap *filtering* adalah tahap pengambilan kata-kata penting dari hasil *tokenizing*. Proses *filtering* dapat menggunakan algoritma *stoplist* (membuang kata-kata yang kurang penting) atau *wordlist* (menyimpan kata penting). *Stoplist* atau *stopwords* adalah kata-kata yang tidak deskriptif yang dapat dibuang (Triawati, 2009).

- *Stemming*

*Stemming* merupakan suatu proses yang terdapat dalam sistem yang mentransformasi kata-kata yang terdapat dalam suatu dokumen ke kata-kata akarnya (*root word*) dengan menggunakan aturan-aturan tertentu (Agusta, 2009).

### 2.1.3. Bag of Words

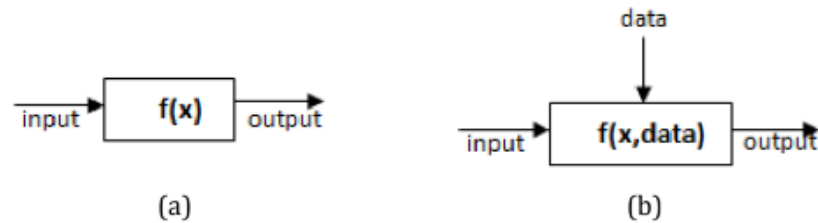
Pendekatan yang paling umum digunakan dalam menyajikan tiap dokumen pada klasifikasi teks adalah pendekatan dengan menggunakan kolom vektor dari jumlah kata. *Bag of words* adalah sebuah objek berupa vektor yang berisi hanya informasi dari jumlah setiap kata. Pada *bag of words*, aturan tata bahasa, ikatan kalimat, dan aturan paragraf diabaikan (Einstein, 2019). Kata-kata pada *bag of words* selanjutnya akan diubah menjadi bentuk vektor agar model dapat memproses data dengan baik.

## 2.2. Machine Learning

*Machine Learning* merupakan suatu aplikasi pada Artificial Intelligence (AI) yang menyediakan sebuah sistem kinerja secara otomatis serta berlatih membenahi diri dari suatu pengalaman atau pengetahuan tanpa diprogram secara eksplisit atau spesifik. *Machine Learning* melibatkan berbagai disiplin ilmu seperti ilmu statistika, ilmu komputer, ilmu matematika, dan bahkan neurologi. Pembelajaran mesin berfokus pada pengembangan program komputer yang bisa mengakses data dan menggunakannya untuk belajar sendiri. Dengan menggunakan data, pembelajaran mesin memungkinkan komputer menemukan wawasan tersembunyi tanpa diprogram secara eksplisit saat mencarinya (Harani & Hasanah, 2020).

*Machine Learning* bermula di awal abad 20, seorang penemu Spanyol, Torres y Quevedo, membuat sebuah mesin *learning* setelah ditemukannya komputer digital. Sebutan *Machine Learning* pada dasarnya merupakan proses komputer untuk berlatih dari data (*learn from data*). Semua pengetahuan dan pemahaman mengenai *Machine Learning* pasti akan selalu melibatkan data (Harani & Hasanah, 2020).

Terdapat perbedaan mendasar antara algoritma *Machine Learning* dengan algoritma pemrograman umum. Pada algoritma pemrograman umum, algoritma ditulis secara eksplisit. Ilustrasi dapat dilihat pada Gambar 2.1 di mana Gambar 2.1a adalah diagram algoritma umum dan Gambar 2.1b adalah diagram algoritma *Machine Learning* (Kusuma, 2012).



Gambar 2.1 Diagram Algoritma Umum dan *Machine Learning* (Kusuma, 2012)

Pada Gambar 2.1a terdapat suatu proses di mana output merupakan luaran fungsi  $f(x)$  di mana  $x$  adalah masukan fungsi. Adapun Gambar 2.1b, output merupakan luaran fungsi  $f$  di mana yang menjadi masukan adalah input dan data.

*Machine Learning* memiliki dua jenis teknik yaitu *Supervised Learning* dan *Unsupervised Learning*. *Supervised Learning* merupakan proses pelatihan model yang telah diketahui masukan (input) dan keluaran (output) datanya. *Unsupervised Learning* merupakan proses pelatihan model dengan cara mendapatkan pola yang tersembunyi atau struktur intrinsik pada data masukan (input) (Harani & Hasanah, 2020).

- *Supervised Learning*

Pembelajaran mesin yang diawasi menghasilkan suatu model yang melatih prediksi bersumber pada bukti adanya ketidakpastian. Algoritma pembelajaran yang diawasi memerlukan seperangkat data masukan dan tanggapan yang diketahui terhadap data (output) dan melatih model untuk menghasilkan prediksi yang masuk akal untuk respons terhadap data baru. *Supervised Learning* biasa digunakan untuk memprediksi keluaran (output) suatu data. Algoritma yang umum digunakan untuk klasifikasi meliputi dukungan mesin vektor (SVM) (Harani & Hasanah, 2020).

- *Unsupervised Learning*

*Unsupervised Learning* menemukan pola tersembunyi dalam suatu data. *Unsuperised Learning* digunakan untuk menarik kesimpulan dari kumpulan data yang terdiri dari data masukan tanpa respons berlabel. *Clustering* adalah teknik belajar tanpa pengamatan yang umum. *Clustering* biasa digunakan untuk melakukan analisis dan eksplorasi dalam menemukan sebuah pola atau pengelompokan terpendam dalam data. Algoritma yang umum menggunakan algoritma clustering meliputi seperti hierarki *clustering*, model campuran Gaussian, *clustering subtractive* dan *clustering fuzzy c-means* (Harani & Hasanah, 2020).

### 2.3. Random Forest Classifier

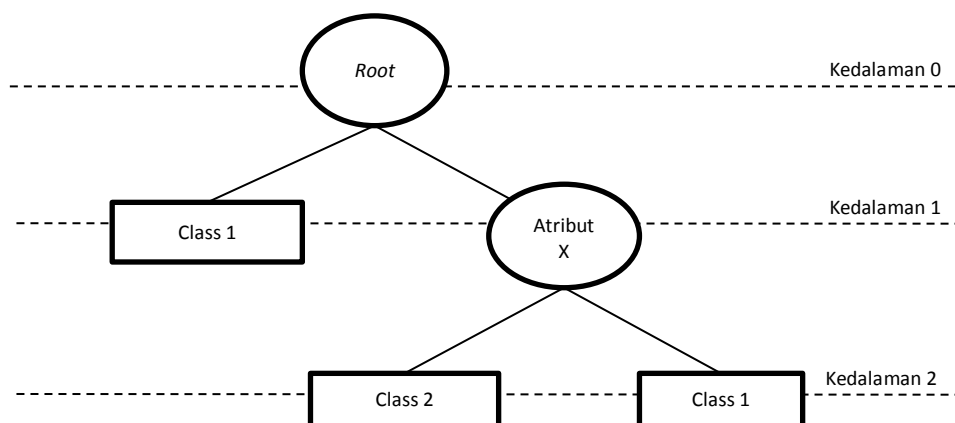
*Random Forest* adalah perangkat berbasis *tree* dengan setiap *tree* bergantung pada koleksi variabel acak. *Random Forest* pertama kali diperkenalkan oleh Leo Breiman yang terinspirasi oleh penelitian sebelumnya oleh Amit dan Geman. *Random Forest* merupakan sebuah penambahan dari ide bagging milik Breiman dan dikembangkan sebagai kompetitor bagi boosting. *Random Forest* dapat digunakan untuk variabel yang membutuhkan respon berupa pengkategorian, yang mengarah pada klasifikasi, atau variabel yang membutuhkan respon berkelanjutan, mengarah pada regresi (Ma & Zhang, 2012).

*Random Forest* terdiri dari banyak *Decision Tree*. Secara umum, *Random Forest* bekerja dengan cara membuat banyak *decision tree* yang digunakan untuk mengklasifikasikan data. Hasil dari masing-masing *tree* tersebut akan *divoting* dan diambil hasil yang paling banyak.

*Random Forest* secara umum menampilkan hasil yang lebih signifikan dibandingkan dengan *single tree classifier* seperti C4.5. Tingkat generalisasi error yang dihasilkan sebanding dengan Adaboost namun lebih tahan terhadap *noise* (Ahmad, Maqsood, Khan, & Jehad, 2012).

#### 2.3.1. Decision Tree

*Tree* adalah graf dengan tipe spesial. *Tree* merupakan sebuah struktur data yang dibentuk dari kumpulan *nodes* dan *edges* disusun dalam sebuah hierarki. *Nodes* dibagi menjadi *internal nodes (split)* dan *terminal nodes (leaf)*. *Internal nodes* dinotasikan sebagai lingkaran, dan terminal dinotasikan sebagai kotak (Criminisi & Shoton, 2013). Contoh *decision tree* dapat dilihat pada Gambar 2.2.



Gambar 2.2 Decision Tree

Pada Gambar 2.2 dapat dilihat bahwa *root* dan Atribut X merupakan *internal nodes* sedangkan *class 1* dan *class 2* merupakan terminal nodes.

Untuk menghasilkan *split* atribut yang optimum, terdapat dua cara penentuan *nodes (criterion)*, yaitu Gini dan Entropy.

#### 2.3.1.1. Gini

Pengukuran *Gini Index* adalah salah satu metode yang digunakan dalam algoritma *decision tree* untuk mendapatkan *split* optimal dari *root* node dan *subsequent splits*. *Gini Index* dihitung dengan formula,

$$GINI_{index} = 1 - \sum_{i=1}^c (p_i)^2 \quad (2.1)$$

Semakin rendah nilai *gini impurity*, semakin tinggi kehomogenan suatu node. Nilai terendah dari *Gini Index* adalah 0, sedangkan nilai tertinggi dari *Gini Index* adalah 0,5. *Gini Index* bernilai 0 jika node sudah menjadi murni, yang berarti bahwa seluruh elemen yang terdapat di dalam node tersebut adalah satu buah kelas yang unik. *Split* yang optimum dipilih dengan melihat nilai *Gini Index* yang paling minimal (Aznar, 2020). Setelah *Gini Index* didapatkan, *Gini Split* akan dihitung dengan menggunakan formula,

$$GINI_{split} = \sum_{i=1}^c \frac{n_i}{n} GINI(i) \quad (2.2)$$

#### 2.3.1.2. Entropy

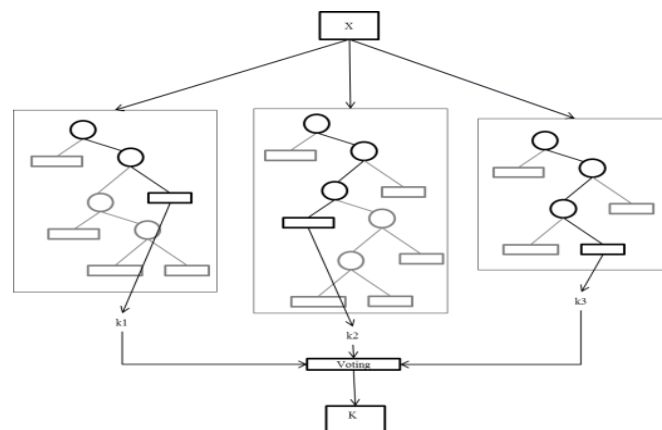
Pengukuran *Information Gain* atau entropi adalah salah satu metode yang digunakan dalam algoritma *decision tree* untuk mendapatkan *split* optimal. Konsep entropi mengukur seberapa informatifnya sebuah *node*. Entropi dihitung dengan formula berikut :

$$Entropy = - \sum_{i=1}^c p_i \log_2(p_i) \quad (2.3)$$

Nilai terendah dari entropi adalah 0, sedangkan nilai tertinggi entropi adalah 1. Entropi akan bernilai 0 saat seluruh elemen yang terdapat di dalam node merupakan satu kelas yang sama. Dalam proses perhitungannya, entropi lebih kompleks karena menggunakan logaritma, sehingga membutuhkan waktu yang lebih lama untuk diproses (Aznar, 2020).

### 2.3.2. Algoritma Random Forest

*Random Forest* adalah perangkat berbasis *tree*. *Random Forest* dapat digunakan untuk pengklasifikasian maupun untuk prediksi regresi. *Random Forest* merupakan pengembangan dari sistem Bagging. Pada bagging, sebagian data dipilih secara acak namun diambil seluruh fiturnya, sedangkan *Random Forest* mengambil sedikit data dan sedikit fitur secara acak untuk kemudian dibuat menjadi beberapa *tree* individual yang tidak berkorelasi (Dangeti, 2017). Contoh bagan *Random Forest* dapat dilihat pada Gambar 2.3.



Gambar 2.3 *Random Forest*

Pada Gambar 2.3 dapat dilihat ide sederhana dari cara kerja *Random Forest Classifier*. Data akan diklasifikasikan oleh *tree* yang berbeda dengan kedalaman yang bervariasi tergantung dari sample yang dipilih. Tiap *tree* akan menghasilkan satu buah hasil klasifikasi, kemudian seluruh hasil klasifikasi tersebut akan di *voting* untuk menentukan *class* terakhir.

*Random Forest* memiliki tiga parameter yang dapat diatur untuk menghasilkan akurasi yang lebih baik dengan beberapa situasi:

- *m*, jumlah variabel prediksi yang dipilih secara acak pada setiap node.
- *J*, jumlah *tree* pada forest.
- *Tree size*, sebagai pengukuran yang dilakukan oleh node terkecil untuk pemisahan atau angka terbesar dari terminal nodes.

Error secara umum dapat semakin berkurang seiring dengan semakin tingginya nilai *J*, namun pada satu titik *J* dapat menjadi terlalu besar sehingga terjadi overfitting. Namun pada *Random Forest*, hal tersebut tidak berlaku. Pada *Random Forest* nilai *J* akan dikonvergenkan ke dalam limit. Nilai *J* dapat dipilih sebesar apapun sesuai keinginan tanpa perlu takut akan menggeneralisasi error (Ma & Zhang, 2012).

Algoritma *Random Forest Classifier* (Ma & Zhang, 2012)

Dimisalkan data *train* sebagai  $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$  dengan  $x_i = (x_{i,1}, \dots, x_{i,p})$ . For  $j = 1$  to  $J$  :

1. Ambil bootstrap sample  $D_j$  sebanyak  $N$  dari  $D$ .
2. Menggunakan bootstrap dari sampel  $D_j$  sebagai data *training*, buat sebuah *tree* dengan binary recursive partitioning dengan langkah :
  - a. Memulai observasi dari satu node
  - b. Mengulangi step berikut untuk setiap nodes hingga bertemu dengan *stopping criterion* :
    - (i) Memilih  $m$  *predictors* secara acak dari  $p$  *predictors* yang tersedia.
    - (ii) Mencari *binary split* terbaik di antara seluruh *binary split* pada  $m$  *predictors* pada langkah (i).
    - (iii) Memecah node menjadi dua node turunan menggunakan *split* dari langkah (ii)

Membuat prediksi menggunakan sistem *voting* dengan rumus klasifikasi  $f(x) = \max_y \sum_{j=1}^J I(h_j(x) = y)$  di mana  $h_j(x)$  adalah hasil prediksi dari respon terhadap  $x$  pada *tree* ke-  $j$ .

Algoritma *Random Forest* dimulai dengan menotasikan data sebagai  $D$  dengan isi data berupa nilai  $x$  dan  $y$  sebanyak  $N$ . Iterasi dilakukan sebanyak  $J$  (jumlah pohon yang akan dibangun). Langkah pertama di dalam iterasi adalah melakukan pengambilan sampel secara acak dari  $D$  sebanyak  $N$ , selanjutnya data akan dibuat menjadi *tree* dengan langkah-langkah sebagai berikut :

- Dimulai dengan mengobservasi seluruh data  $D$  dalam satu *node*.
- Mengulang *step* berikut ini terus menerus hingga *node* tidak dapat dipecah lagi : memilih  $m$  *predictor* secara acak dari  $p$  *predictors* yang tersedia, mencari pembelahan biner terbaik di antara seluruh pembelahan biner dari  $m$  *predictors* menggunakan salah satu dari *criterion* Gini atau Entropy, membelah *nodes* menjadi dua bagian.
- Menentukan hasil prediksi dengan melihat *class* prediksi  $y$  yang paling banyak muncul (*voting*)



## 2.4. Confusion matrix

*Confusion matrix* adalah suatu metode yang biasa digunakan untuk melakukan perhitungan akurasi pada *data mining* atau sistem pendukung keputusan. Pada pengukuran kinerja menggunakan *confusion matrix*, terdapat empat istilah sebagai representasi hasil proses klasifikasi, yaitu *True Positif* (TP), *True Negatif* (TN), *False Positif* (FP) dan *False Negatif* (FN) (Visa, Ramsay, Ralescu, & van der Knaap, 2011). *Confusion matrix* dapat dilihat pada Tabel 2.1.

Tabel 2.1 *confusion matrix*

		True Values	
		True	False
Prediction	True	TP Correct result	FP Unexpected result
	False	FN Missing result	TN Correct absence of result

Nilai *True Negatif* (TN) merupakan jumlah data negatif yang terdeteksi dengan benar, sedangkan *False Positif* (FP) merupakan data negatif namun terdeteksi sebagai data positif. Sementara itu, *True Positif* (TP) merupakan data positif yang terdeteksi benar. *False Negatif* (FN) merupakan kebalikan dari *True Positif*, sehingga data positif, namun terdeteksi sebagai data negatif (Dicoding, 2020).

### 2.4.1. Precision

*Precision* atau presisi merupakan rasio dari prediksi benar positif yang dibandingkan dengan seluruh hasil prediksi positif. Berikut adalah aturan Presisi.

$$Precision = \frac{TP}{(TP + FP)} \times 100\%$$

### 2.4.2. Recall

*Recall* merupakan rasio dari prediksi benar positif yang dibandingkan dengan semua data yang benar positif. Berikut adalah aturan *Recall*.

$$Recall = \frac{TP}{(TP + FN)} \times 100\%$$

### 2.4.3. Akurasi

Akurasi adalah persentase dari total data yang diidentifikasi dan dinilai. Berikut ini adalah aturan akurasi.

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \times 100\%$$

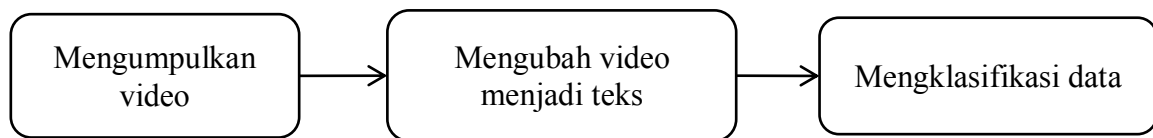
## BAB 3 ANALISIS DAN PERANCANGAN

### 3.1. Analisis Model

Analisis model adalah langkah paling awal yang harus dilakukan dalam perancangan dan pengembangan model. Tujuan dari analisis model adalah untuk mengidentifikasi masalah yang ada pada model, sehingga model dapat berjalan dengan baik. Analisis model dilakukan dengan cara menguraikan komponen-komponen dalam sebuah model. Hal yang perlu dianalisis adalah analisis masalah dan analisis kebutuhan model.

#### 3.1.1. Analisis Masalah

Dalam penelitian ini, permasalahan yang akan dibahas adalah bagaimana cara mengklasifikasikan video ulasan kecantikan yang diunggah ke *platform* YouTube ke dalam bentuk *rating*. Solusi permasalahan tersebut terdapat pada Gambar 3.1.



Gambar 3.1. Bagan Solusi Permasalahan

Secara umum model ini dibagi menjadi tiga bagian, yaitu sebagai berikut :

- Mengumpulkan video, proses ini meliputi pengumpulan data video, penyeleksian video berdasarkan judul, dan pengunduhan video.
- Mengubah video menjadi teks, proses ini meliputi perubahan format video menjadi audio, penerjemahan audio ke dalam bentuk teks, penghapusan baris data yang gagal diterjemahkan, dan pemberian identitas pada setiap baris data.
- Mengklasifikasi data, proses ini meliputi pelatihan dan pengujian data dengan algoritma *Machine Learning Random Forest*.

#### 3.1.2. Analisis Kebutuhan

Analisis kebutuhan dibagi menjadi dua bagian, yaitu analisis kebutuhan fungsional dan kebutuhan nonfungsional. Analisis kebutuhan fungsional menganalisis kebutuhan apa saja yang diperlukan dalam membangun model, sedangkan analisis kebutuhan non-fungsional menganalisis kebutuhan properti apa saja yang digunakan dalam membangun model.

### 3.1.2.1. Analisis Kebutuhan Fungsional

Kebutuhan fungsional merupakan semua proses yang dilakukan oleh model serta menunjukkan fasilitas yang dibutuhkan dalam model. Model diharapkan dapat melakukan fungsi sebagai berikut :

- Model dapat menganalisis video ulasan produk kecantikan dengan baik.
- Model dapat mengeluarkan output berupa *rating* dari data video ulasan produk kecantikan.

### 3.1.2.2. Analisis Kebutuhan Non-Fungsional

Kebutuhan non-fungsional merupakan properti yang dibutuhkan agar model dapat berjalan dengan baik. Berikut aspek-aspek yang dibutuhkan dalam pembuatan model :

#### a) Kebutuhan Perangkat Keras (*Hardware*)

Perangkat keras yang dibutuhkan dalam membangun model adalah satu buah *personal computer*. Semakin tinggi spesifikasi *personal computer* yang digunakan akan semakin baik dan semakin cepat dalam membangun model.

#### b) Kebutuhan Perangkat Lunak (*Software*)

Perangkat lunak yang dibutuhkan dalam membangun model adalah bahasa pemrograman Python dengan library-library yang menunjang pembuatan model. Penggunaan aplikasi berbasis web tambahan seperti Jupyter Notebook atau Google Colab dapat digunakan jika perangkat keras yang digunakan berspesifikasi rendah.

## 3.2. Perancangan Model Analisis Sentimen

### 3.2.1. Pengumpulan Video

Proses pengumpulan video melewati beberapa proses. Proses yang dilakukan dalam pengumpulan video adalah proses *scraping* data, penyeleksian judul video, dan pengunduhan video. Proses pengumpulan video dapat dilihat dalam Gambar 3.2.



Gambar 3.2. Bagan Proses Pengumpulan Video

#### 3.2.1.1. Scraping

*Scraping* dilakukan untuk mengumpulkan data yang akan digunakan sebagai pembelajaran *Machine Learning*. Data yang dikumpulkan berupa kode identitas video, alamat url video, judul video, waktu dan durasi video dari hasil pencarian Youtube

dengan kata kunci pilihan seperti ‘review’, ‘garnier’, ‘somebymi’, dan kata kunci lain untuk mendapatkan url video yang berhubungan dengan ulasan produk kecantikan. Hasil scraping data adalah berupa file berekstensi .csv seperti pada Tabel 3.1.

Tabel 3.1. Hasil *Scraping* Data

youID	url	title	Duration
_OLLISE4S-g	<a href="https://www.youtube.com/watch?v=_OLLISE4S-g">https://www.youtube.com/watch?v=_OLLISE4S-g</a>	Some By Mi Miracle Serum Review 30 Hari Pemakaian!	PT6M58S
fQ6g1_yr1iU	<a href="https://www.youtube.com/watch?v=fQ6g1_yr1iU">https://www.youtube.com/watch?v=fQ6g1_yr1iU</a>	REVIEW JUJUR PEMAHAIAN SOME BY MI AHA BHA PHA 30 DAYS MIRACLE TONER & MIRACLE SERUM SECARA RUTIN	PT7M44S
NF4hS1wJCxE	<a href="https://www.youtube.com/watch?v=Nf4hS1wJCxE">https://www.youtube.com/watch?v=Nf4hS1wJCxE</a>	SERUM UNTUK JERAWAT?? SOME BY MI AHA BHA PHA 30 DAYS MIRACLE SERUM REVIEW - Novie Maru	PT5M57S
jdcbYF1ovwE	<a href="https://www.youtube.com/watch?v=jdcbYF1ovwE">https://www.youtube.com/watch?v=jdcbYF1ovwE</a>	Review & Testing Skincare Korea SOME BY MI + First Impression ðŸ˜ƒ	PT7M2S

### 3.2.1.2. Penyeleksian Judul Video

Data yang telah dikumpulkan selanjutnya diseleksi sehingga menghasilkan sebuah data dengan format *comma separated values* (csv) yang berisi alamat url video sesuai dengan yang diharapkan. Ketentuan yang digunakan dalam menyeleksi judul video adalah sebagai berikut :

- Durasi video minimal 1 menit dan maksimal 12 menit 59 detik.
- Judul video berisi hanya satu jenis produk (hanya salah satu dari *toner*, *serum*, *cream*, atau *facial wash*).
- Judul video tidak mengandung kata yang menunjukkan banyak jenis produk seperti ‘kit’, ‘rangkaiannya’, ‘full’, ‘one brand’, dan ‘products’.
- Judul video berisi hanya satu merk dagang (tidak memiliki kata, ‘battle’, ‘perbandingan’, dan ‘X’).
- Judul video tidak mengandung kata ‘cara’, ‘how to’, ‘rekomendasi’, ‘pilih’ untuk menghindari video yang bukan berisi ulasan.

Setelah diseleksi, kolom selain kolom url akan dihapus untuk persiapan mengunduh video seperti pada Tabel 3.2 .

Tabel 3.2. Hasil Penyeleksian Judul Video

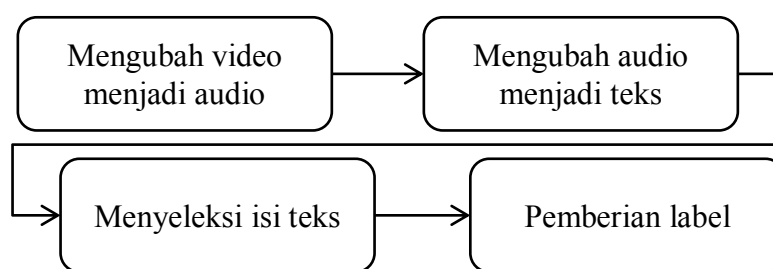
url
<a href="https://www.youtube.com/watch?v=_OLLISE4S-g">https://www.youtube.com/watch?v=_OLLISE4S-g</a>
<a href="https://www.youtube.com/watch?v=NF4hS1wJCxE">https://www.youtube.com/watch?v=NF4hS1wJCxE</a>
<a href="https://www.youtube.com/watch?v=R43bWbDcE90">https://www.youtube.com/watch?v=R43bWbDcE90</a>
<a href="https://www.youtube.com/watch?v=ruSia5qfRb0">https://www.youtube.com/watch?v=ruSia5qfRb0</a>

### 3.2.1.3. Pengunduhan Video

Data hasil penyeleksian judul video selanjutnya akan diunduh. Video akan diunduh dengan bantuan *library* Pytube. Nama asli video akan diubah dengan format ‘video\_angka’ untuk memudahkan proses selanjutnya. Video diunduh dengan format mp4 tanpa gambar untuk mengecilkan ukuran video. Kumpulan video tersebut disimpan dalam satu buah folder.

### 3.2.2. Pengubahan Video Menjadi Teks

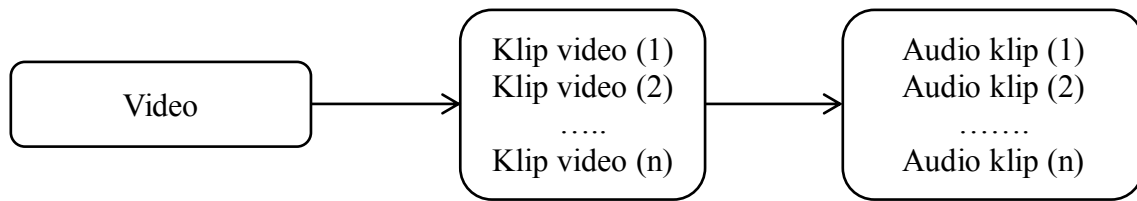
Pengubahan video menjadi teks melalui beberapa proses. Video yang telah diunduh akan diubah menjadi audio berekstensi wav, selanjutnya audio tersebut akan diubah ke dalam bentuk teks. Data teks tersebut kemudian akan diberikan identitas (labelisasi) agar data dapat dilatih. Proses pengubahan video menjadi teks dapat dilihat pada Gambar 3.3.



Gambar 3.3. Bagan Proses Perubahan Video Menjadi Teks

#### 3.2.2.1. Video menjadi Audio

Proses pengubahan video menjadi audio berekstensi wav sangat dibutuhkan agar selanjutnya dapat diubah ke dalam bentuk teks. Proses pengubahan video berekstensi mp4 menjadi audio berekstensi wav dilakukan dengan bantuan *library* moviepy. Proses pengubahan dapat dilihat pada Gambar 3.4.



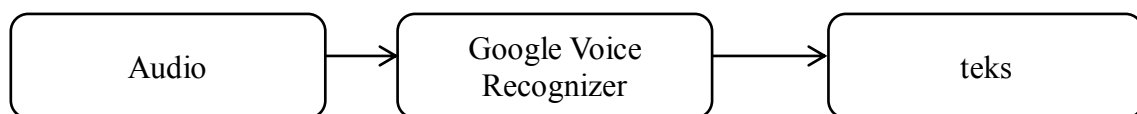
Gambar 3.4. Bagan Proses Pengubahan Video Menjadi Audio

Proses pengubahan video menjadi audio adalah sebagai berikut :

- Video dipecah menjadi beberapa klip video dengan durasi masing-masing maksimal dua menit
- Masing-masing klip video akan diubah menjadi audio dengan ekstensi wav.
- Klip audio tersebut kemudian disimpan dalam folder.

#### 3.2.2.2. Audio menjadi Teks

Proses mengubah video menjadi teks dibantu oleh library Speech\_Recognizer. Library tersebut menggunakan Google Speech Recognition untuk mengubah audio berekstensi wav menjadi tulisan. Proses pengubahan audio menjadi teks dapat dilihat pada Gambar 3.5.



Gambar 3.5 Bagan Proses Mengubah Audio Menjadi Teks

Hasil rekognasi tersebut kemudian akan disimpan dalam bentuk csv. Hasil rekognasi tersebut dapat dilihat dalam Tabel 3.3.

Tabel 3.3 Contoh Hasil Rekognasi Audio menjadi Teks

id	file_name	Text
0	video_0.mp4	Toner dari Garnier ini bagus banget mencerahkan dan juga melembabkan banget pokoknya bagus
1	Video_1.mp4	Aku pakai serum dari Some by Mi hijau ini ternyata gak cocok kulitku beruntusan dan terasa tipis banget
2	Video_2.mp4	Baru pertama kali aku coba pakai krim malam dari ponds hasilnya bagus bener-bener melembabkan tapi gak mencerahkan di kulit aku

#### 3.2.2.3. Penyeleksian Isi Teks

Data teks tersebut kemudian akan diseleksi. Data yang akan dihapus adalah data yang di dalam kolom teks mengandung hal seperti berikut :

- Kolom teks kosong (gagal direkognasi).
- Kolom teks berisi jumlah huruf kurang dari 200 huruf
- Kolom teks mengandung kata '*battle*' di dalamnya.

#### 3.2.2.4. Pemberian Identitas (Labelisasi)

Proses pemberian identitas dilakukan secara manual untuk meningkatkan akurasi pelabelan tersebut. Proses labelisasi bertujuan untuk memberikan identitas pada tiap ulasan sebagai data latih model. Contoh hasil labelisasi dapat dilihat dalam Tabel 3.4.

Tabel 3.4. Contoh Hasil Labelisasi

id	file_name	Teks	label
1	video_36.mp4	Produk serum some by mi yang botol hijau ini bikin wajah aku beruntusan banget dan banyak banget jerawatnya	1
2	video_82.mp4	Pembersih wajah garnier yang warna hijau ini rasanya kering banget kulit jadi terasa kaku tapi kelihatan bersih banget	2
3	video_51.mp4	Krim malam dari ponds ini beneran melembabkan tapi aku gak ngerasain efek mencerahkannya	3
4	video_129.mp4	Serum garnier yang kuning ini bener-bener mencerahkan walaupun bikin timbul beruntusan sedikit tapi kulit rasanya segar banget	4
5	Video_111.mp4	Acnes facial foam ini enak banget dipakenya gak bikin muka kering lembab banget dan baunya juga harum banget recommended pokoknya	5

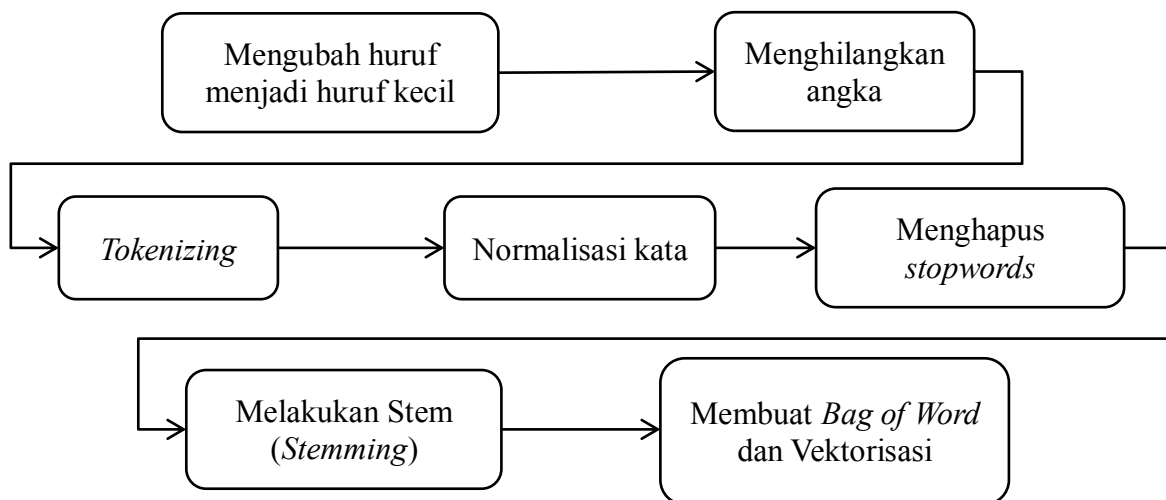
Label berupa angka 1 sampai 5. Ulasan negatif diberi angka 1, ulasan agak negatif diberi angka 2, ulasan netral diberi angka 3, ulasan agak positif diberi angka 4, dan ulasan positif diberi angka 5.

#### 3.2.3. Pengklasifikasian Data

Data yang telah diberi label akan dilatih agar dapat mengklasifikasikan data dengan baik. Proses yang dilalui untuk mengklasifikasi data melalui tiga tahap, yaitu tahap *Preprocessing*, tahap pembagian data, dan tahap pelatihan data.

### 3.2.3.1. *Preprocessing*

*Preprocessing* merupakan tahap yang sangat penting dalam pengklasifikasian data. Data dengan teks yang sangat panjang akan diproses melalui beberapa tahap sehingga menghasilkan sebuah teks yang lebih singkat, padat, dan jelas, namun tidak mengurangi makna teks tersebut. *Preprocessing* memiliki beberapa tahapan. Tahapan pada *preprocessing* seperti pada Gambar 3.6.



Gambar 3.6. Bagan Tahap *Preprocessing*

Berdasarkan Gambar 3.6, *preprocessing* terdiri dari tujuh tahapan. Ketujuh tahapan yang dilakukan dalam *Preprocessing* adalah sebagai berikut :

a) Mengubah huruf menjadi huruf kecil

Semua huruf akan diubah menjadi huruf kecil untuk menyeragamkan bentuk huruf.

b) Menghilangkan angka

Semua angka akan dihilangkan untuk mengurangi *noise* pada data.

c) *Tokenizing*

Kalimat akan dipecah menjadi kumpulan kata untuk mempermudah proses berikutnya.

d) Normalisasi kata

Kata-kata yang tidak baku atau kata-kata yang mengalami kesalahan pada saat tahap perekognasiannya sehingga kata yang terdapat pada teks tersebut tidak sesuai dengan kata yang dimaksud oleh pembicara akan dinormalisasi sehingga menghasilkan kata-kata yang lebih rasional dan lebih sesuai dengan maksud pembicara. Contoh kata yang akan dinormalisasi dapat dilihat dalam Tabel 3.5.



Tabel 3.5. Contoh Normalisasi Kata

No.	Kata Asli	Kata setelah Normalisasi
1	Certain	Sertakan
2	Pengen	Ingin
3	Ingredients	Komposisi
4	Ngobatin	Mengobati
5	Gara-gara	Karena

e) Menghapus *stopwords*

*Stopwords* adalah kata-kata umum yang biasa muncul dalam jumlah besar dan dianggap tidak memiliki makna. Terdapat tiga *stopwords* yang digunakan pada penelitian, ketiga *stopwords* itu adalah *stopwords* yang telah terdaftar dalam *library* Sastrawi, *stopwords* bahasa Indonesia yang telah terdaftar dalam *library* NLTK, dan *stopwords* tambahan yang dibuat sendiri. Contoh kata-kata yang merupakan *stopwords* dapat dilihat dalam Tabel 3.6.

Tabel 3.6. Contoh Stopwords

Nya	yang	doang	Deh	tuh
Si	sih	Nih	Hai	halo
Aja	untuk	Nah	Ya	loh

f) Melakukan stem (*stemming*)

*Stemming* adalah tahap mengembalikan sebuah kata kembali ke kata dasarnya dengan menghilangkan semua imbuhan. Proses *stemming* dibantu oleh *stemmer* pada *library* Sastrawi. Contoh kata-kata yang distem dapat dilihat dalam Tabel 3.7.

Tabel 3.7 Contoh Stem Kata

No	Kata Asli	Kata Setelah Distem
1	Mencari	Cari
2	Pengembangan	Kembang
3	Mengakibatkan	Akibat
4	Mencoba-coba	Coba
5	Mengobati	Obat

g) Membuat *Bag of Words* dan Vektorisasi

Kantong kata dibuat dengan mengumpulkan data yang telah diproses sebelumnya ke dalam matriks. Matriks yang dibuat terdiri dari dua variabel, yaitu  $x$  untuk menampung matriks dari kolom teks yang telah dibersihkan dan  $y$  untuk menampung matriks dari kolom *rating*. Matriks yang dibuat dapat dilihat pada Tabel 3.8.

Tabel 3.8 Contoh Matriks *Bag of Words*

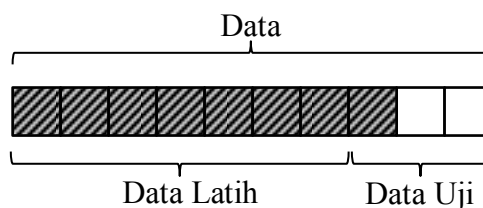
Data	bagus	Wajah	cerah	Jerawat	hilang	suka	banget	lembab	kulit
bagus wajah cerah	1	1	1	0	0	0	0	0	0
jerawat hilang	0	0	0	1	1	0	0	0	0
suka banget	0	0	0	0	0	1	1	0	0
wajah lembab jerawat hilang	0	1	0	1	1	0	0	1	0
kulit cerah bagus banget	1	0	1	0	0	0	1	0	1

Apabila data memiliki kata yang sama dengan kolom, kotak (data,kolom) akan berisi angka 1. Jika data tidak memiliki kata yang sama dengan kolom, kotak (data,kolom) akan berisi angka 0. Jumlah kata di dalam kantong adalah sebanyak kata yang terdapat dalam keseluruhan data.

### 3.2.3.2. Pembagian Data

Data dibagi menjadi dua bagian, yaitu data latih untuk melatih model, dan data uji untuk menguji model. Perbandingan yang digunakan untuk membagi data latih dan

data uji adalah 8:2. Perbandingan pembagian data menjadi data latih dan data uji dapat dilihat pada Gambar 3.7.



Gambar 3.7. Perbandingan Data Uji dan Data Latih

a) Data Latih

Data latih memiliki perbandingan lebih banyak dari data uji. Hal ini bertujuan untuk meningkatkan akurasi pemrosesan data. Data latih berisi satu buah matriks data dengan satu kolom yang lebih banyak dibanding data uji. Hal ini dikarenakan data latih memiliki kolom hasil yang menampung *rating* data.

b) Data Uji

Data uji memiliki perbandingan yang lebih sedikit dari data latih. Data uji memiliki isi yang sama dengan data latih, namun data uji memiliki satu kolom lebih sedikit dibanding data latih. Hal ini dikarenakan data uji belum memiliki kolom hasil yang menampung *rating* data.

### 3.2.3.3. Pelatihan Data

Pelatihan data adalah proses utama pengklasifikasian ulasan sehingga suatu data ulasan akan menghasilkan prediksi *rating* ulasan. Pelatihan data menggunakan *Machine Learning* dengan metode pengklasifikasian *Random Forest Classifier*. Dengan menggunakan algoritma *Random Forest Classifier*, data akan dibentuk menjadi beberapa *Decision Tree* dengan jumlah atribut dan kedalaman yang bervariasi.

Jumlah *tree* yang digunakan adalah jumlah *tree* dengan nilai akurasi paling tinggi di antara 2 hingga 100 *tree*. Penentuan jumlah *tree* akan dilakukan dengan beberapa kali percobaan. Percobaan pertama dimulai dengan memeriksa akurasi model dengan jumlah *tree* 10, 20, 30, 40, 50, 60, 70, 80, 90, 100. Percobaan akan dilanjutkan dengan jumlah *tree* di selang angka tersebut.

*Tree* akan dibuat dengan kriteria Gini, yaitu menghitung dan mengambil atribut dengan tingkat ketidakmurnian yang paling rendah. Nilai ketidakmurnian masing-masing atribut akan dihitung dengan formula *Gini Index* (2.1) dan *Gini split* (2.2). Pembuatan *tree* dilakukan dengan mengambil sampel dengan jumlah acak dan atribut

dengan jumlah acak. Sampel dan atribut yang digunakan sebagai contoh pembuatan satu buah *tree* dapat dilihat pada Tabel 3.9.

Tabel 3.9 Sampel data membuat *tree*

	Cerah	jerawat	cocok	class
wajah jadi kusam kering tidak cocok banget	0	0	1	5
kulit cerah banget	1	0	0	1
cocok ini	0	0	1	1
jerawat parah banget tidak cocok	0	1	1	5
cocok banget kulit tidak jerawat	0	1	1	1
kulit jadi cerah cocok banget	1	0	1	1
kulit jadi jerawat tapi cerah lembab	1	1	0	2
jerawat hilang senang banget	0	1	0	1
jerawat hilang cerah agak kering	1	1	0	2
bikin lembab banget cocok di wajah	0	0	1	1

Dari sampel diketahui atribut yang akan dijadikan sebagai acuan untuk membuat *tree* adalah atribut cerah, jerawat, dan cocok. Dalam 10 sampel terdapat 3 *class*, yaitu *class* 1 yang menampung ulasan positif, *class* 2 yang menampung ulasan agak positif, dan *class* 5 yang menampung ulasan negatif.

Untuk mengetahui atribut apa yang optimal digunakan sebagai pembelah *root*, perhitungan Gini Index dan Gini *Split* dibutuhkan. Perhitungan Gini Index dan *Split* dapat dilihat pada Tabel 3.10.

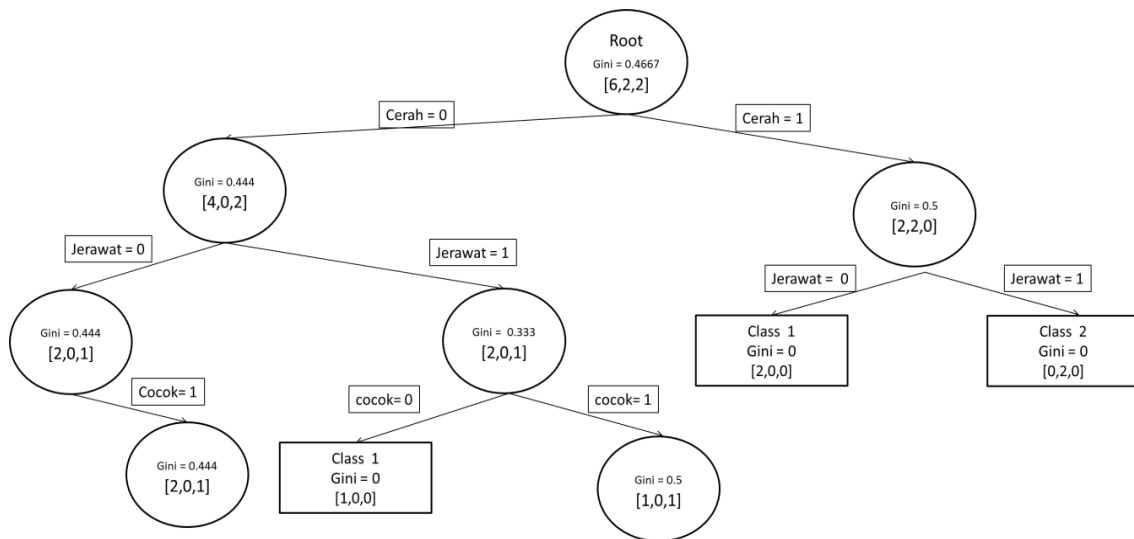
Tabel 3.10 Perhitungan Gini Index dan Gini *Split*

Atribut		jumlah	class 1	class 2	class 5	Gini	
Total		10	6	2	2	Index	Split
Cerah	0	6	4	0	2	0.444444	0.44667
	1	4	2	2	0	0.5	
Jerawatan	0	5	4	0	1	0.32	1.2
	1	5	2	2	1	0.64	
Cocok	0	4	2	2	0	0.5	0.933333
	1	6	4	0	2	0.444444	

Dari hasil perhitungan didapatkan atribut cerah sebagai atribut dengan indeks *split* yang paling kecil. Bentuk *root tree* akan menjadi seperti pada Gambar 3.8.

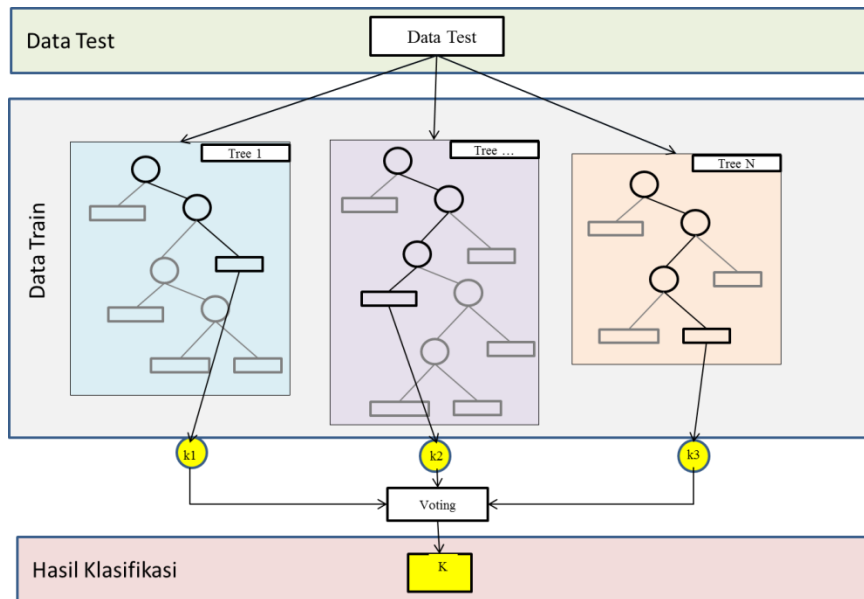
Gambar 3.8 Pembentukan *root tree*

Tahap penghitungan gini index dan gini *split* terus dilakukan di setiap cabang, hingga gini pada terminal bernilai 0 atau hingga atribut tidak bisa dibagi lagi. Hasil akhir dari *tree* sampel dapat dilihat pada Gambar 3.9.

Gambar 3.9 pembentukan *tree sample*

Jumlah *tree* yang akan dibuat tergantung dari hasil percobaan pada penentuan jumlah *tree*. Atribut dan sampel data akan dipilih secara acak untuk membangun lebih dari satu *tree*.

Setelah kumpulan *tree* berhasil dibuat, model *Random Forest Classifier* akan terlihat seperti pada Gambar 3.10.



Gambar 3.10 Alur *Random Forest*

Data Latih akan digunakan untuk membuat *tree* dan melatih model. Data uji yang masuk nantinya akan diklasifikasikan oleh masing-masing *tree* dengan kombinasi atribut yang berbeda. Hasil klasifikasi dari masing-masing *tree* kemudian akan *divoting* dan diambil hasil klasifikasi terbanyak .

## **BAB 4**

### **IMPLEMENTASI DAN PENGUJIAN**

#### **4.1. Implementasi**

##### **4.1.1. Lingkungan Eksperimen**

Perangkat yang digunakan untuk membangun model analisis sentimen pada video ulasan produk kecantikan menggunakan *Random Forest Classifier* terbagi menjadi perangkat keras dan perangkat lunak.

##### **4.1.1.1. Perangkat Keras**

Perangkat keras yang digunakan dalam membangun model analisis sentimen pada video ulasan produk kecantikan menggunakan *Random Forest Classifier* adalah sebuah PC (*Personal Computer*) dengan keterangan sebagai berikut :

- *Windows Edition* : Windows 7 Ultimate
- *Manufacturer* : ASUSTeK COMPUTER INC.
- *Processor* : Intel(R) Celeron(R) CPU 1007U @ 1.50GHz (2 CPUs), ~1.5Ghz
- *RAM* : 4.00 GB
- *System Type* : 64-bit Operating System,

##### **4.1.1.2. Perangkat Lunak**

Perangkat lunak yang digunakan dalam model analisis sentimen pada video ulasan produk kecantikan menggunakan *Random Forest Classifier* adalah sebagai berikut :

##### **a) Google Colaboratory**

*Google colab* adalah *source code editor* berbasis *cloud* yang mendukung hampir semua *library* yang dibutuhkan dalam lingkungan pengembangan *Artificial Intelligence (AI)*. Dengan spesifikasi GPU Nvidia K80s, T4s, P4s, dan P100s, RAM 13 GB, Disk 180 GB.

##### **b) Python v.3.7.0.**

Python adalah bahasa pemrograman interpretatif multiguna yang memakai filosofi perancangan dengan fokus kepada tingkat keterbacaan kode. Sebagai bahasa pemrograman, Python menggabungkan kemampuan, kapabilitas dan sintaksis kode serta fungsi pustaka yang berkualitas tinggi.

Library python yang digunakan dalam analisis sentimen pada video ulasan produk kecantikan menggunakan *Random Forest Classifier* adalah

- Pandas v.1.0.3
- Moviepy v.1.0.3
- Pytube v.10.4.1
- Speech\_recognition v.3.8.1
- NLTK v.3.2.5
- Re v.2.2.1
- Sklearn v.0.22.2.
- PySastrawi v.1.2.0
- Seaborn v.0.11.1.
- Matplotlib v.3.2.2.

#### 4.1.2. Dataset

Dataset yang digunakan dalam pembuatan model dibuat dengan cara mengumpulkan data video, mengunduh video, merekognasi audio pada video ulasan ke dalam bentuk teks, menyeleksi data, dan memberikan label secara manual pada teks ulasan. Setelah melalui tahapan-tahapan tersebut, terkumpul satu buah dataset berformat *comma separated values* (csv).

Hasil akhir dari pembentukan dataset menghasilkan sebuah dataset yang terdiri dari 59 data dengan distribusi *class* seperti dalam Tabel 4.1.

Tabel 4.1 Distribusi *Class* Dataset

<i>Rating</i>	<i>Class</i>	Jumlah Data
5	1	28
4	2	18
3	3	7
2	4	3
1	5	3
Total data		59

Dari Tabel 4.1 dapat diketahui bahwa persebaran dataset berpusat pada *class* 1 dan *class* 2. Dataset berisi tiga kolom, yaitu kolom *file\_name*, *text*, dan *rating*. Kolom *file\_name* berisi nama dari file video, kolom *text* berisi hasil rekognasi audio dari video ulasan, dan kolom *rating* berisi *rating* dari ulasan tersebut.

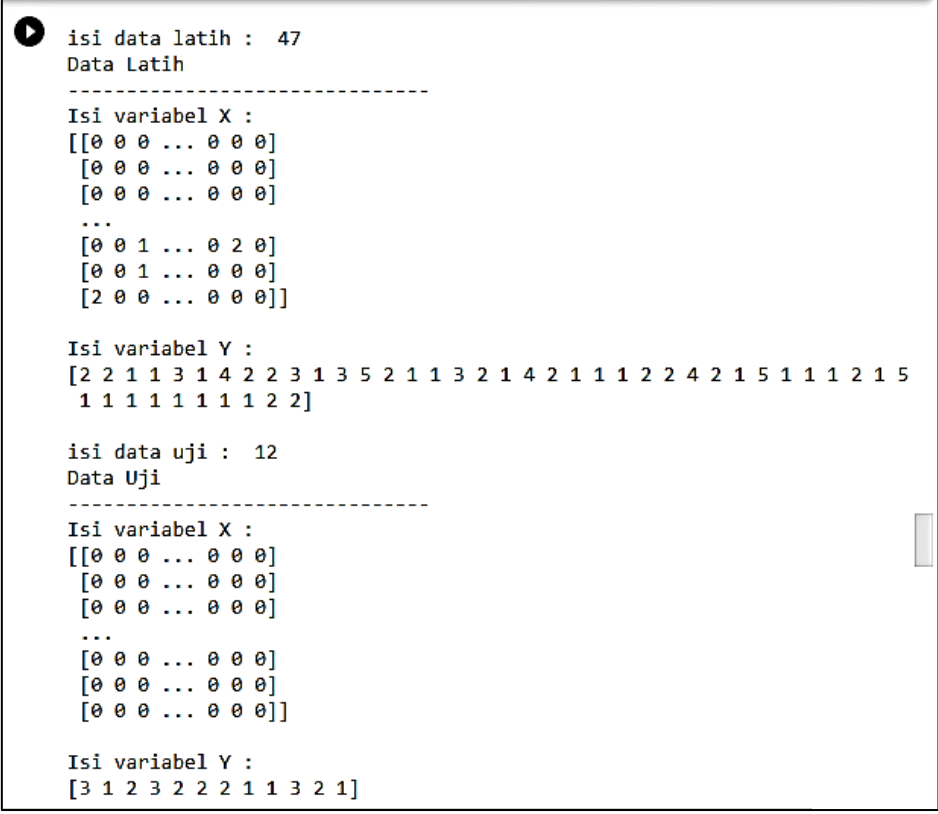




Setelah dilakukan vektorisasi, variabel X berisi matriks dengan dimensi (59 , 2574). Variabel X berisi 59 baris data dan 2574 kolom kata. Variabel Y berisi matriks dengan dimensi (59 , 1). Variabel Y berisi 78 baris data dan 1 kolom hasil.

#### 4.1.3.2. Pembagian data

Data dibagi dengan perbandingan data latih dan data uji sebesar 8:2. Hasil pembagian data dapat dilihat pada Gambar 4.4.



```
isi data latih : 47
Data Latih
-----
Isi variabel X :
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 1 ... 0 2 0]
 [0 0 1 ... 0 0 0]
 [2 0 0 ... 0 0 0]]

Isi variabel Y :
[2 2 1 1 3 1 4 2 2 3 1 3 5 2 1 1 3 2 1 4 2 1 1 1 2 2 4 2 1 5 1 1 1 2 1 5
 1 1 1 1 1 1 1 1 2 2]

isi data uji : 12
Data Uji
-----
Isi variabel X :
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]

Isi variabel Y :
[3 1 2 3 2 2 2 1 1 3 2 1]
```

Gambar 4.4 Pembagian Data Latih dan Data Uji

Dari 59 data dibagi menjadi data latih dan data uji dengan perbandingan 8:2. Data latih berisi 47 data sedangkan data uji berisi 12 data.

#### 4.1.3.3. Pengklasifikasian dengan *Random Forest*

Metode yang digunakan adalah *Random Forest Classifier*. Proses pembangunan model dibantu oleh *Random Forest Classifier* milik *library* sklearn.

Dalam memilih jumlah *tree* untuk proses klasifikasi melalui beberapa percobaan untuk menghasilkan akurasi yang paling tinggi. Percobaan pertama mencoba untuk mengklasifikasikan data dengan jumlah pohon 10, 20, 30, 40, 50, 60, 70, 80, 90, dan 100. Hasil pengecekan akurasi percobaan pertama dapat dilihat pada Tabel 4.2.

Tabel 4.2 Percobaan akurasi jumlah *tree* pertama

10	20	30	40	50	60	70	80	90	100
0.416	0.416	0.416	0.416	0.416	0.416	0.416	0.416	0.416	0.416

Dari percobaan pertama dapat dilihat bahwa pada jumlah *tree* 10, 20, 30, 40, 50, 60, 70, 80, 90, dan 100 memiliki nilai akurasi yang sama. Pengecekan akan dilanjutkan dengan mengecek akurasi pada jumlah *tree* 5, 15, 25, 35, 45, 55, 65, 75, 85, dan 95. Hasil pengecekan akurasi percobaan kedua dapat dilihat pada Tabel 4.3.

Tabel 4.3 Percobaan akurasi jumlah *tree* kedua

5	15	25	35	45	55	65	75	85	95
0.5	0.333	0.416	0.416	0.416	0.416	0.416	0.416	0.416	0.416

Dari percobaan kedua dapat dilihat bahwa jumlah *tree* 5 memiliki nilai akurasi tertinggi. Nilai akurasi kemudian menurun pada jumlah *tree* 15. Pada jumlah *tree* 25 nilai akurasi kembali naik.

Dari percobaan kedua dapat dilihat bahwa nilai akurasi tertinggi berada pada jumlah 5 *trees* dengan nilai akurasi sebesar 0.5. Jumlah *trees* yang akan digunakan dalam model klasifikasi *Random Forest* analisis sentiment video ulasan produk kecantikan menggunakan *Random Forest* adalah sebanyak 5 *trees*.

Hasil dari pengklasifikasian menggunakan *Random Forest Classifier* dengan 5 *trees* dapat dilihat pada Gambar 4.5.

```

Isi variabel y_test :
[3 1 2 3 2 2 2 1 1 3 2 1]

Isi variabel y_pred :
[1 1 1 1 3 2 1 1 1 1 2 1]

```

Gambar 4.5 Hasil Klasifikasi Model

*Confusion matrix* yang terbentuk dari hasil pengklasifikasian model dapat dilihat pada Tabel 4.4.

Tabel 4.4 *Confusion matrix* Pengujian

		Prediksi		
		class1	class 2	class 3
Nilai Asli	class 1	4	0	0
	class 2	2	2	1
	class3	3	0	0

Dari *confusion matrix* dapat dilihat bahwa dari 12 data, model dapat memprediksi 6 data secara tepat. Error yang terjadi pada model adalah kesalahan memprediksi *class 2* sebagai *class 1* (2 data), *class 2* sebagai *class 3* (1 data), dan *class 3* sebagai *class 1* (3 data).

Nilai akurasi dihitung dari tabel *confusion matrix* menggunakan rumus akurasi adalah sebagai berikut :

$$\begin{aligned}
 \text{Akurasi} &= \frac{TP}{\text{Total Data}} \times 100\% \\
 \text{Akurasi} &= \frac{(4) + (2) + (0)}{12} \times 100\% \\
 \text{Akurasi} &= \frac{6}{12} \times 100\% \\
 \text{Akurasi} &= 50\%
 \end{aligned}$$

Setelah dihitung menggunakan rumus akurasi didapatkan nilai akurasi pada model yang telah dibuat adalah sebesar 50%. Hasil perhitungan menandakan bahwa terdapat 50% data yang diprediksi secara tepat.

Nilai presisi dihitung dari tabel *confusion matrix* menggunakan rumus presisi sebagai berikut :

$$\begin{aligned}
 \text{Presisi} &= \frac{TP}{TP + FP} \times 100\% \\
 P(\text{class 1}) &= \frac{4}{4 + 5} = \frac{4}{9} = 0.444 \\
 P(\text{class 2}) &= \frac{2}{2 + 0} = \frac{2}{2} = 1 \\
 P(\text{class 3}) &= \frac{0}{0 + 1} = \frac{0}{1} = 0 \\
 \text{Semua Presisi} &= \frac{P(\text{class1}) + P(\text{class2}) + P(\text{class3})}{\text{Jumlah class}} \times 100\% \\
 \text{Semua Presisi} &= \frac{(0.444) + (1) + (0)}{3} \times 100\% \\
 \text{Semua Presisi} &= \frac{1.444}{3} \times 100\% \\
 \text{Semua Presisi} &= 48.1\%
 \end{aligned}$$

Setelah dihitung menggunakan rumus akurasi didapatkan nilai akurasi pada model yang telah dibuat adalah sebesar 48.1%. Hasil perhitungan menandakan bahwa terdapat 48.1% data yang diprediksi benar dari keseluruhan data yang diprediksi sebagai ulasan positif.

Nilai *recall* dihitung dari tabel *confusion matrix* menggunakan rumus *recall* sebagai berikut :

$$\begin{aligned}
 Recall &= \frac{TP}{TP + FN} \times 100\% \\
 R(class\ 1) &= \frac{4}{4 + 0} = \frac{4}{4} = 1 \\
 R(class\ 2) &= \frac{2}{2 + 3} = \frac{2}{5} = 0.4 \\
 R(class\ 3) &= \frac{0}{0 + 3} = \frac{0}{3} = 0 \\
 \\ 
 Semua\ Recall &= \frac{R(class1) + R(class2) + R(class3)}{Jumlah\ class} \times 100\% \\
 Semua\ Recall &= \frac{(1) + (0.4) + (0)}{3} \times 100\% \\
 Semua\ Recall &= \frac{1.40}{3} \times 100\% \\
 Semua\ Recall &= 46.7\%
 \end{aligned}$$

Setelah dihitung menggunakan rumus akurasi didapatkan nilai akurasi pada model yang telah dibuat adalah sebesar 46.7%. Hasil perhitungan menandakan bahwa terdapat 46.7% data yang diprediksi benar dibandingkan dengan data yang diprediksi sebagai ulasan negatif.

#### 4.2. Pengujian

Model yang telah diimplementasi akan diuji coba. Dari dataset yang telah diberi label, diambil 5 data yang mengulas produk kecantikan dengan merk Garnier. Pengujian dilakukan pada 3 produk Garnier yaitu :

- Garnier Sakura White Facial Foam
- Garnier Sakura White Booster Serum
- Garnier Vitamin C Booster Serum

Distribusi *class* pada pengujian ulasan 5 produk kecantikan merk Garnier adalah sebagai berikut :

- Ulasan dengan *rating* 5 sebanyak 1 data
- Ulasan dengan *rating* 4 sebanyak 3 data
- Ulasan dengan *rating* 1 sebanyak 1 data

Hasil pengujian model pada 5 ulasan produk kecantikan dengan merk Garnier dapat dilihat pada Tabel 4.5.

Tabel 4.5 Pengujian Model

No	Nama File	Nama Produk	Rangkuman Teks	Rating asli	Rating Prediksi	Keterangan
1	Video_77	Garnier Sakura White Facial Foam	<ul style="list-style-type: none"> <li>• Cocok untuk kulit sensitif</li> <li>• Bikin kulit glowing dan cerah</li> </ul>	5	5	Berhasil
2	Video_115	Garnier Sakura White Booster Serum	<ul style="list-style-type: none"> <li>• Cepat meresap ke kulit</li> <li>• Kulit terlihat makin cerah dan segar</li> <li>• Terasa lembab</li> <li>• Bekas jerawat tidak hilang setelah pemakaian seminggu</li> </ul>	4	4	Berhasil
3	Video_124	Garnier Sakura White serum	<ul style="list-style-type: none"> <li>• Wajah lumayan lebih cerah</li> <li>• Wajah jadi lebih lembab</li> <li>• Belum bisa memudarkan bekas jerawat</li> </ul>	4	4	Berhasil
4	Video_134	Garnier Skin Natural Sakura White Serum	<ul style="list-style-type: none"> <li>• Lama menyerap di kulit berminyak</li> <li>• Tidak mendapat khasiat setelah seminggu pemakaian</li> </ul>	1	4	Gagal

			<ul style="list-style-type: none"> <li>• Timbul jerawat di dagu setelah pemakaian</li> </ul>			
5	Video_123	Garnier Vitamin C Booster Serum	<ul style="list-style-type: none"> <li>• Wajah lumayan cerah</li> <li>• Kulit wajah terlihat lebih segar</li> </ul>	4	4	Berhasil

### 4.3. Analisis

Setelah melewati tahapan pembentukan model dan pengujian, analisis dilakukan terhadap dua hal, yaitu analisis berdasarkan studi kasus dan analisis pada model.

#### 4.3.1. Analisis Studi Kasus

*Random Forest Classifier* dapat digunakan untuk memprediksi rating pada video ulasan. Hasil dari analisis sentimen ulasan produk kecantikan dapat digunakan sebagai tolok ukur dalam mengetahui seberapa positif atau negatif ulasan yang diberikan oleh *Beauty Vlogger* yang mengupload video ulasan suatu produk kecantikan pada platform YouTube. Rating ulasan produk kecantikan tersebut diharapkan dapat digunakan sebagai penunjang pengambilan keputusan bagi produsen produk kecantikan.

#### 4.3.2. Analisis Model

Model dapat memprediksi ulasan dengan akurasi sebesar 50%. Terdapat beberapa faktor yang mungkin akan memengaruhi nilai akurasi model, yaitu :

- Kurang meratanya distribusi *class* data sehingga besar kemungkinan terjadi kesalahan prediksi pada *class* data yang lebih sedikit.
- Teks ulasan diambil dan direkognasi langsung dari video sehingga terdapat kata yang tidak sesuai dengan maksud si pembicara.
- Penggunaan kata sehari-hari yang tidak dikenali oleh *stemmer* dan *stopwords remover* membuat *noise* pada teks menjadi besar.
- Kurangnya isi dari list normalisasi dan list stopwords menyebabkan masih adanya *noise*.
- Penggunaan seluruh bagian video menyebabkan *claim* produk juga ikut terproses sehingga memengaruhi hasil klasifikasi menjadi kurang akurat.

## **BAB 5 PENUTUP**

### **5.1. Kesimpulan**

Berdasarkan hasil yang di dapat dari seluruh rangkaian penelitian ini diperoleh simpulan bahwa algoritma *Random Forest Classifier* dapat digunakan untuk melakukan analisis sentimen pada video ulasan produk kecantikan. Algoritma *Random Forest* berhasil mengklasifikasi video ulasan produk kecantikan ke dalam 5 *class rating* menggunakan 59 data video dengan data latih sebanyak 80% (47 data) dan data uji sebanyak 20% (12 data). Berdasarkan model yang telah dibuat klasifikasi berhasil dilakukan dengan jumlah *tree* sebanyak 5 dengan tingkat akurasi sebesar 50%.

### **5.2. Saran**

Penelitian tugas akhir ini belum sempurna, ada beberapa hal yang dapat diperbaiki dan dikembangkan dari penelitian ini. Beberapa hal yang dapat dilakukan untuk memperbaiki dan mengembangkan model ini adalah sebagai berikut :

- Menggunakan data yang lebih banyak.
- Menggunakan komputer dengan spesifikasi yang lebih baik.
- Menggunakan model klasifikasi yang berbeda dengan tingkat akurasi yang lebih baik.
- Memperbanyak list normalisasi dan list *stopwords* untuk mengurangi *error*.
- Meratakan distribusi *class* pada dataset agar akurasi model bisa lebih baik lagi.



## DAFTAR REFERENSI

- Dicoding*. (2020, Agustus 4). Retrieved Desember 3, 2020, from <https://www.dicoding.com/blog/machine-learning-adalah/>
- Ahmad, N., Maqsood, I., Khan, R., & Jehad, A. (2012). Random Forest and Decision Trees. *IJCSI International Journal of Computer Science*, 9(5).
- Aznar, P. (2020, 12 2). *quantdare.com*. (Quantdare) Retrieved 2 14, 2021, from <https://quantdare.com/decision-trees-gini-vs-entropy/>
- Bhandary, U. (2019). Detection of Hate Speech in Videos using Machine Learning. *SJSU Scholar Works*.
- Criminisi, A., & Shotton, J. (2013). *Decision Forests for Computer Vision and Medical Image Analysis*. New York: Springer London Heidelberg New York Dordrecht.
- Dangeti, P. (2017). *Statistic for Machine Learning*. Birmingham: Pakct Publishing.
- Einstein, J. (2019). *Introduction of Natural Language Processing*. Cambridge: The Massachusetts Institute of Technology.
- Harani, N. H., & Hasanah, M. (2020). *Deteksi Objek dan Pengenalan Karakter Plat Nomor Kendaraan Indonesia Berbasis Python*. Bandung: Kreatif Industri Nusantara.
- Kusuma, P. D. (2012). *Machine Learning Teori, Program, dan Studi Kasus*. Sleman: Penerbit Deepublish.
- Ma, Y., & Zhang, C. (2012). *Ensamble Machine Learning*. New York: Springer New York Dordrecht Heidelberg London.
- Mariezka, F. I., Haufar, H., & Yustikasari. (2018). Pemaknaan Profesi Beauty Vlogger melalui Pengalaman Komunikasi. *Nyimak Journal of Communication*, 2, 95-111.
- MonkeyLearn*. (n.d.). (MonkeyLearn) Retrieved February 1, 2021, from <https://monkeylearn.com/sentiment-analysis/>
- Ruben, W. (2020, Mei 20). *Medium*. Retrieved Januari 13, 2021, from <https://medium.com/prinsip-dasar-natural-language-processing/prinsip-dasar-natural-language-processing-f415d5d48af3>
- Visa, S., Ramsay, B., Ralescu, A., & van der Knaap, E. (2011). Confusion Matrix-based Feature Selection. *Midwest Artificial intelligence and Cognitive Science Conference*, 7(April).