**IMPLEMENTASI ALGORITMA RANDOM FOREST MENGGUNAKAN WORD EMBEDDING UNTUK ANALISIS SENTIMEN DENGAN PENERAPAN TRANSFER LEARNING**

**SKRIPSI**

**Diajukan sebagai salah satu syarat untuk memperoleh gelar**

**Sarjana Komputer (S.Kom.)**

****

**Ahsanul Qalbi Fajar Islami**

**00000015655**

**PROGRAM STUDI INFORMATIKA**

**FAKULTAS TEKNIK DAN INFORMATIKA**

**UNIVERSITAS MULTIMEDIA NUSANTARA**

**TANGERANG**

**2020**

# **HALAMAN PERSETUJUAN**

**Skripsi dengan judul**

**IMPLEMENTASI ALGORITMA RANDOM FOREST MENGGUNAKAN WORD EMBEDDING UNTUK ANALISIS SENTIMEN DENGAN PENERAPAN TRANSFER LEARNING**

oleh

Ahsanul Qalbi Fajar Islami

00000015655

**telah disetujui untuk diajukan pada**

**Sidang Ujian Skripsi Universitas Multimedia Nusantara**

Tangerang, 5 April 2020

Menyetujui,

Dosen Pembimbing 1 Dosen Pembimbing 2

Julio Christian Young, S.Kom., M.Kom. Andre Rusli, S.Kom., M.Sc.

Ketua Program Studi Informatika

Nunik Afriliana, S.Kom., MMSI

# **PERNYATAAN TIDAK MELAKUKAN PLAGIAT**

Dengan ini saya:

Nama : Ahsanul Qalbi Fajar Islami

NIM : 00000015655

Program Studi : Informatika

Fakultas : Teknik dan Informatika

Menyatakan bahwa Skripsi yang berjudul “**Implementasi Algoritma Random Forest Menggunakan Word Embedding Untuk Analisis Sentimen Dengan Penerapan Transfer Learning**” ini adalah karya ilmiah saya sendiri, bukan plagiat dari karya ilmiah yang ditulis oleh orang lain atau lembaga lain, dan semua karya ilmiah orang lain atau lembaga lain yang dirujuk dalam Skripsi ini telah disebutkan sumber kutipannya serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/ penyimpangan, baik dalam pelaksanaan Skripsi maupun dalam penulisan laporan Skripsi, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk mata kuliah Skripsi yang telah saya tempuh.

Tangerang, 05 April 2020

Ahsanul Qalbi Fajar Islami

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

Sebagai sivitas akademik Universitas Multimedia Nusantara, saya yang bertanda tangan di bawah ini:

Nama : Ahsanul Qalbi Fajar Islami

NIM : 00000015655

Program Studi : Informatika

Fakultas : Teknik dan Informatika

Jenis Karya : Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui dan memberikan izin kepada **Universitas Multimedia Nusantara** hak Bebas Royalti Non-eksklusif (*Non-exclusive Royalty-Free Right*) atas karya ilmiah saya yang berjudul:

**Implementasi Algoritma Random Forest Menggunakan Word Embedding Untuk Analisis Sentimen Dengan Penerapan Transfer Learning**

beserta perangkat yang diperlukan.

Dengan Hak Bebas Royalti Non-eksklusif ini, pihak **Universtias Multimedia Nusantara** berhak menyimpan, mengalihmedia atau *format*-kan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mendistribusi dan menampilkan atau mempublikasikan karya ilmiah saya di internet atau media lain untuk kepentingan akademis, tanpa perlu meminta izin dari saya maupun memberikan royalty kepada saya, selama tetap mencantumkan nama saya sebagai penulis karya ilmiah tersebut.

Demikian pernyataan ini saya buat dengan sebenarnya untuk dipergunakan sebagaimana mestinya.

Tangerang, 05 April 2020

Ahsanul Qalbi Fajar Islami

# **HALAMAN PERSEMBAHAN/ MOTO**

*[INSERT QUOTES]*

# **KATA PENGANTAR**

Puji syukur kepada Tuhan Yang Maha Esa karena atas berkat dan rahmat-Nya, penulis dapat menyelesaikan skripsi dan laporan skripsi yang berjudul “Implementasi Algoritma Random Forest Menggunakan Word Embedding Untuk Analisis Sentimen Dengan Penerapan Transfer Learning” dengan tepat waktu.

Proses penyelesaian laporan skripsi tidak terlepas dari bantuan berbagai pihak. Penulis mengucapkan terima kasih kepada:

1. Dr. Ninok Leksono, Rektor Universitas Multimedia Nusantara, yang memberi inspirasi bagi penulis untuk berprestasi,
2. Dr. Eng. Niki Prastomo, S.T., M.Sc. Dekan Fakultas Teknik dan Informatika Universitas Multimedia Nusantara,
3. Nunik Afriliana, S.Kom., MMSI., Ketua Program Studi Informatika Universitas Multimedia, yang menerima penulis dengan baik untuk berkonsultasi,
4. Julio Christian Young, S.Kom., M.Kom. dan Andre Rusli, S.Kom., M.Sc. yang membimbing pembuatan skripsi dan yang telah mengajar penulis tata cara menulis karya ilmiah dengan benar.
5. Ayah dan Ibu, serta keluarga besar yang telah memberikan dukungan moral dan spiritual selama proses penulisan skripsi,
6. Sahabat-sahabat mahasiswa terkasih yang telah membantu memberikan dukungan langsung maupun tidak langsung, bahkan ada yang menemani penulis selama proses penulisan skripsi.

Semoga laporan skripsi ini dapat bermanfaat, baik sebagai sumber informasi maupun sumber inspirasi, bagi para pembaca.

Tangerang, 05 April 2020

Ahsanul Qalbi Fajar Islami

**IMPLEMENTASI ALGORITMA RANDOM FOREST MENGGUNAKAN WORD EMBEDDING UNTUK ANALISIS SENTIMEN DENGAN PENERAPAN TRANSFER LEARNING**

# **ABSTRAK**

Kata Kunci: -----------------------------------------------------------------------------

**RANDOM FOREST ALGORITHM IMPLEMENTATION USING WORD EMBEDDING FOR SENTIMENT ANALYSIS WITH TRANSFERS LEARNING APPLICATION**

# **ABSTRACT**

Keywords: -----------------------------------------------------------------------------

# **DAFTAR ISI**

[**HALAMAN PERSETUJUAN** ii](#_Toc35598713)

[**PERNYATAAN TIDAK MELAKUKAN PLAGIAT** iii](#_Toc35598714)

[**PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS** iv](#_Toc35598715)

[**HALAMAN PERSEMBAHAN/ MOTO** v](#_Toc35598716)

[**KATA PENGANTAR** vi](#_Toc35598717)

[**ABSTRAK** viii](#_Toc35598718)

[**ABSTRACT** ix](#_Toc35598719)

[**DAFTAR ISI** x](#_Toc35598720)

[**DAFTAR GAMBAR** xii](#_Toc35598721)

[**DAFTAR TABEL** xiii](#_Toc35598722)

[**PENDAHULUAN** 14](#_Toc35598723)

[**1.1** **Latar Belakang Masalah** 14](#_Toc35598724)

[**1.2** **Rumusan Masalah** 19](#_Toc35598725)

[**1.3** **Batasan Masalah** 20](#_Toc35598726)

[**1.4** **Tujuan Penelitian** 20](#_Toc35598727)

[**1.5** **Manfaat Penelitian** 20](#_Toc35598728)

[**LANDASAN TEORI** 22](#_Toc35598729)

[**2.1** **Rekayasa Kebutuhan** 22](#_Toc35598730)

[**2.2** **User Feedback** 24](#_Toc35598731)

[**2.3** **Analisa Sentimen** 25](#_Toc35598732)

[**2.4** **Word Embedding** 25](#_Toc35598733)

[**2.5** **FastText** 27](#_Toc35598734)

[**2.6** **Decision Tree** 29](#_Toc35598735)

[**2.7** **Random Forest Classifier** 33](#_Toc35598736)

[**METODOLOGI PENELITIAN DAN PERANCANGAN SISTEM** 35](#_Toc35598737)

[**3.1** **Metodologi Penelitian** 35](#_Toc35598738)

[**3.1.1.** **Studi Literatur** 35](#_Toc35598739)

[**3.1.2.** **Pengumpulan Data** 35](#_Toc35598740)

[**3.1.3.** **Prosedur Penelitian** 35](#_Toc35598741)

[**3.1.4.** **Penulisan Laporan** 38](#_Toc35598742)

[**3.2** **Perancangan Aplikasi** 38](#_Toc35598743)

[**3.2.1** **Flowchart Proses Klasifikasi** 39](#_Toc35598744)

[**3.2.1** **Flowchart Deployed Version** 43](#_Toc35598745)

[**IMPLEMENTASI DAN ANALISIS** 48](#_Toc35598746)

[**4.1** **Spesifikasi Sistem** 48](#_Toc35598747)

[**4.2** **Implementasi Sistem** 48](#_Toc35598748)

[**6.2.1** **Hasil Implementasi CNN** 48](#_Toc35598749)

[**6.2.2** **Hasil Tampilan Aplikasi** 48](#_Toc35598750)

[**SIMPULAN DAN SARAN** 49](#_Toc35598751)

[**5.1** **Simpulan** 49](#_Toc35598752)

[**5.2** **Saran** 49](#_Toc35598753)

[**DAFTAR PUSTAKA** 50](#_Toc35598754)

[**DAFTAR RIWAYAT HIDUP** 53](#_Toc35598755)

# **DAFTAR GAMBAR**

[Gambar 3. 1 39](#_Toc35599633)

# **DAFTAR TABEL**

[Tabel 3. 1 Struktur Tabel Data 46](#_Toc35599669)

**BAB** **I**

# **PENDAHULUAN**

## **Latar Belakang Masalah**

Analisis sentimen adalah salah satu bidang studi yang menganalisis pendapat, sentimen, evaluasi, penilaian, sikap, dan emosi seseorang terhadap suatu entitas tertentu seperti, jasa, organisasi, individu, masalah, peristiwa, topik dan atribut lainnya (Liu, 2012). Analisis sentimen dapat dimanfaatkan untuk melakukan ekstraksi opini-opini dari dokumen, komentar, sosial media, *review blog*, dan data-data lainnya. Analisis sentimen memliki beberapa metode, salah satu diantaranya adalah metode pendekatan *machine learning*. Analisis sentimen dengan pendekatan *machine learning* dapat menggunakan teknik klasifikasi.

Teknik klasifikasi adalah teknik dalam *data mining* untuk mengelompokkan data berdasarkan keterikatan data terhadap data sampel (Irvi Oktanisa, dan Afif Supianto 2018). Salah satu teknik klasifikasi adalah Random Forest. Random Forest adalah salah satu metode berbasis klasifikasi dan regresi dimana terdapat proses agregasi pohon keputusan (Dhawangkhara, dan Riksakomara 2017). Kelebihan dari algoritma Random Forest diantaranya adalah dapat menghindari *overfitting,* meminimalisir waktu *training data*, berjalan secara efisien pada data yang banyak, dan dapat mempertahankan akurasi walaupun sebagian data hilang.

Penelitian terkait mengenai klasifikasi teks dengan algoritma Random Forest dan TF-IDF telah banyak dilakukan salah satunya adalah penelitian oleh M Ali

Fauzi. (2018) yang mengklasifikasikan teks Bahasa Indonesia dengan algoritma Random Forest menggunakan fitur Bag-of-word dan metode pembobotan Term Frequency Inverse Document Fequency (TF.IDF). Disimpulkan bahwa penelitian tersebut memberikan hasil akurasi yang baik yaitu kinerja dengan skor out-of-bag (OOB) 0,829. Penelitian terkait dengan penggunaan TF-IDF selanjutnya adalah penelitian yang dilakukan oleh Yulius Denny dkk (2019) yang menerapkan algoritma TF-IDF untuk *text mining*. Penelitian tersebut menyimpulkan bahwa penggunaan TF-IDF merupakan metode yang tepat untuk digunakan dalam pencarian kata ditiap dokemen dan dapat membantu pengguna mendapatkan dokumen terkait sesuai dengan *query* yang telah di inputkan.

Penelitian terkait mengenai analisis sentimen dengan menggunakan teknik N-gram telah banyak dilakukan, salah satunya adalah penelitian oleh Wahyu Candra Inddhiarta (2017) yang melakukan analisis sentimen pemilihan kepala daerah Jakarta dengan menggunakan N-gram dan algoritma Naïve Bayes. Dari penelitian tersebut, disimpulkan bahwa berdasarkan ketiga jenis token N-gram yaitu unigram, bigram, dan trigram dengan metode Naïve Bayes, nilai akurasi tertinggi terdapat pada penggunaan bigram yaitu 0,823, menunjukan bahwa dengan menggunakan bigram ketepatan akurasi dari sistem lebih baik dari pada unigram dan trigram. Nilai presisi tertinggi juga terdapat pada penggunaan bigram yaitu 0,76. Namun pada nilai *recall* nilai tertinggi terdapat pada penggunaan trigram yaitu 0,898, sehingga disimpulkan penggunaan bigram dalam pengklasifikasian data lebih baik daripada menggunakan unigram atau trigram

Analisis sentimen dengan menggunakan *machine learning*, diperlukan *datasets* sebagai data *training* dan data *testing*. Metode pendekatan *machine learning* menghasilkan akurasi yang baik, akan tetapi performa dari klasifikasinya bergantung pada dataset yang digunakan untuk data *training* masalah ini berkaitan dengan *transfer learning. Transfer learning* adalah metode Deep Learning yang menerapkan pengetahuan atau *knowledge* dari domain yang berbeda namun terkait ke domain tujuan (Di Zhang dkk, 2019). Berdasarkan penelitian yang telah ada sebelumnya, penelitian ini mencoba untuk menganalisis seberapa besar pengaruh *transfer learning* terhadap akurasi klasifikasi analisis sentimen dengan menggunakan algoritma Random Forest dengan bantuan N-gram dan TF-IDF.

## **Rumusan Masalah**

Berdasarkan latar belakang yang telah dijabatkan sebelumnya, rumusan masalah dalam penelitian ini adalah sebagai berikut.

1. Bagaimana mengimplementasikan algoritma Random Forest dengan N- gram dan TF-IDF pada analisis sentimen?
2. Seberapa besar tingkat performa model yang dirancang dengan menggunakan metode *transfer learning* ?

## **Batasan Masalah**

Batasan masalah dalam penelitian ini dapat dijabarkan menjadi beberapa poin sebagai berikut.

1. *Datasets* terdiri dari kumpulan *review* dari pengguna pada produk yang dimiliki oleh perusahaan Amazon, Yelp, dan IMDb.
2. *Datasets review* yang digunakan berupa teks Bahasa Inggris
3. *Transfer learning* akan dilakukan dengan mengirimkan pengetahuan dari satu model Random Forest Classifier berupa fitur-fitur dan nilai kepentingan setiap fitur-fitur tersebut kepada model baru yang akan dibangun, sehingga model tersebut tidak menggunakan fitur yang dianggap tidak penting.

## **Tujuan Penelitian**

Berdasarkan rumusan masalah yang dijelaskan sebelumnya, penelitian ini memiliki tujuan sebagai berikut.

1. Menerapkan algoritma Random Forest dengan menggunakan metode Bigrams dan TF-IDF untuk klasifikasi analisis sentimen *review* oleh *user*.
2. Mengukur dan mengetahui seberapa besar hasil performa model jika diterapkan metode *transfer learning* pada analisis sentimen dengan algoritma Random Forest.

## **Manfaat Penelitian**

Hasil dari penelitian ini diharapkan dapat menerapkan metode *transfer learning* dengan algortima Random Forest pada analisis sentimen sehingga jika memberi pengaruh yang positif untuk performa model diharapkan dapat menanggulangi keterbatasan set data untuk membangun model yang serupa.

.

**BAB II**

# **LANDASAN TEORI**

**2.1 Analisis Sentimen**

Sentimen analisis adalah riset komputasional dari opini sentimen dan emosi yang diekspresikan secara tekstual (Ira Zulfa, dan Edi Winarko 2017). Dalam analisis sentimen, teks data yang didapatkan akan diklasifikasikan menjadi beberapa jenis seperti teks sentimen “positif”, “negatif”, dan “netral”. Pada penerapannya, analisis sentimen dimanfaatkan untuk memberikan nilai reputasi pada pelayanan pelanggan, produk perusahaan, dan reputasi seorang tokoh publik.

**2.2 Word Embedding**

*Word Embedding* adalah istilah yang digunakan untuk teknik mengubah sebuah kata menjadi sebuah *vector* atau *array* yang terdiri dari kumpulan angka. *Word Embedding* adalah sebuah pendekatan yang digunakan untuk merepresentasikan *vector* kata. *Word Embedding* merupakan pengembangan komputasi permodelan kata-kata yang sederhana seperti perhitungan menggunakan jumlah dan frekuensi kemunculan kata dalam sebuah dokumen (Yulius Denny dkk, 2019)

Contoh cara tradisional untuk membaca teks dan mengubah menjadi vektor angka, misalnya terdapat sebuah kalimat yakni “sore ini merupakan sore yang indah”. Langkah pertama adalah membuat sebuah *dictionary* yang berisi *list* dari seluruh kata yang *unique* atau tidak berulang, sehingga *dictionary* yang terbentuk adalah [“Sore”, “ini”, “merupakan”, “yang”, “indah”]. Langkah selanjutnya adalah menggunakan metode *one-shot encoding* yang akan mengeluarkan *output* vektor berupa vektor ‘1’ merepresentasikan tempat kata tersebut pada *list*, dan vektor ‘0’ untuk merepresentasikan tempat kata lainnya. Contoh vektor representasi pada kata ‘merupakan’ mengacu pada metode *one-shot encoding* adalah [0, 0, 1, 0 ,0].

**2.3 Teknik N-gram**

Menurut Wahyu Candra Indhiarta (2017) N-gram merupakan penggabungan kata sifat yang sering muncul untuk menunjukan suatu sentimen. Teknik N-gram memiliki jenis-jenisnya berupa Unigram (n = 1), Bigram (n = 2), Trigram (n = 3), dan seterusnya. Berikut merupakan contoh penerapan N-gram pada kalimat “Pembelajaran mesin merupakan salah satu mata kuliah jurusan informatika”.

Tabel 1. Jenis-jenis N-Gram

|  |  |
| --- | --- |
| **N-Gram** | **Hasil Penerapan** |
| Unigram | Pembelajaran, mesin, merupakan, salah, satu, mata, kuliah, jurusan, informatika |
| Bigram | Pembelajaran mesin, mesin merupakan, merupakan salah, salah satu, satu mata, mata kuliah, kuliah jurusan, jurusan informatika |
| Trigram | Pembelajaran mesin merupakan, mesin merupakan salah, merupakan salah satu, salah satu mata, satu mata kuliah, mata kuliah jurusan, kuliah jurusan informatika |

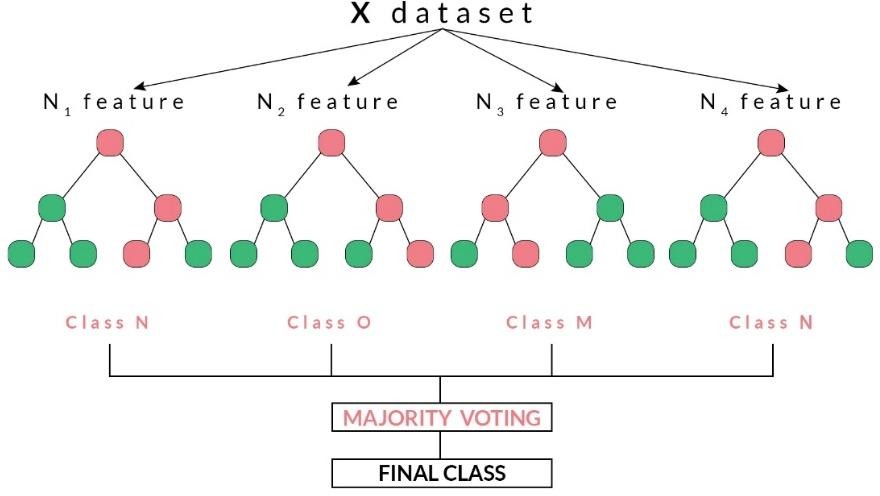
**2.4 Term Frequency-Inverse Document Frequency (TF-IDF)**

Term Frequency-Inverse Document Frequency atau TF-IDF merupakan algoritma yang berguna untuk mengetahui bobot setiap kata atau seberapa sering kata tersebut. Musfiroh Nurjannah, dkk. (2013) menyatakan bahwa metode ini menggabungkan perhitungan bobot, yaitu frekuensi kemunculan sebuah kata di dalam sebuah dokumen (Term Frequency) tertentu dan *inverse* frekuensi dokumen yang mengandung kata tersebut (Inverse Document Frequency).

Frekuensi kemunculan kata (Term Frequency) di dalam dokumen menunjukkan seberapa penting kata tersebut di dalam dokumen tersebut. Frekuensi dokumen yang mengandung kata tersebut (Inverse Document Frequency) menunjukkan seberapa umum kata tersebut. Sehingga bobot hubungan antara sebuah kata dan sebuah dokumen akan tinggi apabila frekuensi kata tersebut tinggi di dalam dokumen dan frekuensi keseluruhan dokumen yang mengandung kata tersebut yang rendah pada kumpulan dokumen Musfiroh Nurjannah, dkk. (2013).

**2.5 Random Forest Classifier**

Random Forest adalah salah satu teknik *machine learning* yang dapat digunakan untuk melakukan klasifikasi. Random Forest merupakan salah satu metode dalam *decision tree*. Menurut (Aditya Yanuar, 2018) *decision tree* atau pohon pengambil keputusan adalah sebuah diagram alir yang berbentuk seperti pohon yang memiliki sebuah *root node* yang digunakan untuk mengumpulkan data, Sebuah *inner node* yang berada pada *root node* yang berisi tentang pertanyaan tentang data dan sebuah *leaf node* yang digunakan untuk memecahkan masalah serta membuat keputusan. Random Forest memiliki beberapa *decision tree*, kemudian algoritma Random Forest mengambil keputusan berdasarkan hasil *voting* terbanyak dari semua *decision tree*. Kelebihan dari Random Forest adalah jika terdapat data yang hilang dengan jumlah tertentu, Random Forest masih dapat melakukan klasifikasi dengan akurasi yang stabil karena tidak bergantung dengan satu *decision tree* saja melainkan membandingkan data *voting decision tree* lainnya.

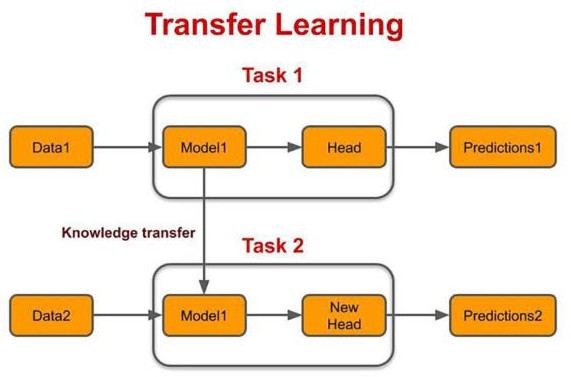


Gambar 1 Struktur Algortima Random Forest

(Shagufta, 2019)

**2.6 Transfer Learning**

*Transfer learning* adalah metode Deep Learning yang menerapkan pengetahuan atau *knowledge* dari domain yang berbeda namun terkait ke domain tujuan (Di Zhang dkk, 2019). Menurut (Reza Fuad, 2018) *transfer learning* bertujuan untuk mengurangi penggunaan set data dengan skala besar. Pada penelitian ini *transfer learning* akan dilakukan dengan membangun satu model dasar Random Forest Classifier dan mengirimkan pengetahuan dari model tersebut berupa *feature importance* kepada model baru yang akan dibangun, sehingga tidak menggunakan fitur yang dianggap tidak penting.



Gambar 2 Visualisasi gambaran *transfer learning*

(Pratik, 2019)

**BAB III**

# **METODOLOGI PENELITIAN DAN PERANCANGAN SISTEM**

## **Metodologi Penelitian**

### **Studi Literatur**

Membaca dan mempelajari sumber dari jurnal ilmiah dan karya tulis terkait dengan topik yang diteliti adalah yang dilakukan selama tahap studi literatur.

### **Pengumpulan Data**

Pengumpulan datasets dilakukan dengan menggunakan *datasets riview* dari

*user* pada produk perusahaan dengan Bahasa Inggris.

### **Prosedur Penelitian**

Penelitian akan dilakukan dalam beberapa tahap sebagai berikut.

1. *Import Libraries*

Sebelum melakukan proses klasifikasi lebih lanjut, *library* yang dibutuhkan akan di-*import* terlebih dahulu. *Library* yang di-*import* tersebut akan mendukung proses penelitian agar lebih efektif.

### **Penulisan Laporan**

Setelah Random Forest untuk klasifikasi sentimen sekiranya berhasil dilakukan, maka hasil uji coba akan dilakukan pencatatan dan dokumentasi ke dalam bentuk laporan.

## **Perancangan Aplikasi**

### **Flowchart Proses Klasifikasi**

### **Flowchart Deployed Version**

**3.2.3** **Rancangan Struktur Tabel**

### **Rancangan Tampilan Antarmuka**

**BAB IV**

# **IMPLEMENTASI DAN ANALISIS**

## **Spesifikasi Sistem**

## **Implementasi Sistem**

### **Hasil Implementasi CNN**

### **Hasil Tampilan Aplikasi**

**BAB V**

# **SIMPULAN DAN SARAN**

## **Simpulan**

## **Saran**

# **DAFTAR PUSTAKA**

# **DAFTAR RIWAYAT HIDUP**