PROPOSAL SKRIPSI

**DETEKSI BERITA PALSU DI TWITTER MENGGUNAKAN NAÏVE BAYES DAN ADABOOST**



Disusun oleh :

Wisnu Anggoro Putro

11190940000028

PROGRAM STUDI MATEMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI SYARIF HIDAYATULLAH

JAKARTA

2022 M / 1443 H

PROPOSAL SKRIPSI

**DETEKSI BERITA PALSU DI TWITTER MENGGUNAKAN NAÏVE BAYES DAN ADABOOST**

Diajukan oleh :

**Wisnu Anggoro Putro**

11190940000028

Usulan Dosen Pembimbing

|  |  |
| --- | --- |
| Dosen Pembimbing I  <Nama Dosen>  <NIP XXXXXX>  Bio Singkat Dosen Pembimbing I (1 paragraf) |  |
| Dosen Pembimbing II  <Nama Dosen>  <NIP XXXXXX>  Bio Singkat Dosen Pembimbing II (1 paragraf) |  |

# Latar belakang

## Rumusan Masalah

Perkembangan teknologi informasi membuat pertukaran informasi menjadi semakin mudah. Munculnya media sosial sepert *Twitter, Facebook, Instagram, Google, dan Youtube* telah merubah cara pandang masyarakat dalam mengakses berita. Media sosial merupakan salah satu media komunikasi yang sangat populer ini. Berdasarkan data yang dirilis oleh Databoks pada Januari 2022 *Twitter* memiliki 18,45 juta pengguna yang berasal dari Indonesia.

Rumusan masalah dari penelitian ini adalah sebagai berikut:

1. Bagaimana mengimplementasikan metode *Multilayer Perceptron* untuk menganalisis sentimen pada ulasan aplikasi Discord di Google Playstore?

2. Bagaimana Discord meningkatkan kualitas layanannya berdasarkan ulasan dari Google Playstore?

## Manfaat dan Tujuan Penelitian

Penelitian ini diharapkan dapat bermanfaat bagi masyarakat agar tidak salah dalam memilih produk *skincare*, serta agar Scarlett Whitening dapat selalu menjaga kualitas produknya sehingga kepercayaan masyarakat tetap terjaga.

Tujuan penulisan penelitian ini adalah:

1. Mengimplementasikan metode *Naïve Bayes* dan *Adaboost* untuk mendeteksi *fake reviews* produk Scarlett Whitening di aplikasi Twitter.

2. Mengukur performa dari metode *Naïve Bayes* dan *Adaboost* untuk mendeteksi *fake revews* pada produk Scarlett Whitening di aplikasi Twitter.

# TINJAUAN PUSTAKA DAN LANDASAN TEORI

## Data Mining

Data Mining adalah suatu tahap ekplorasi dari “deteksi informasi dalam file” prosedur untuk menentukan desain dalam koleksi data besar termasuk pendekatan terhadap hubungan simulasi kecerdasan buata, Machine Learning, dan perekaman sistem. Selain itu, Data Mining dapat diartikan suatu keterampilan baru yang berlaku untuk mendukung yang menekankan pada materi yang lebih penting dalam penyimpanan data mereka [21].

Data Mining merupakan teknik mengektraksi pengetahuan secara implisit yang berguna dari data yang memiliki jumlah cukup besar. Data Mining telah banyak digunakan untuk menyelesaikan masalah bisnis seperti segmentasi pelanggan, retensi pelanggan, penilaian kredit, rekomendasi prduk, pemasaran langsung, penjualan silang, deteksi penipuan, dan sebagainya [23].

Data mining atau penambangan data melibatkan berbagai fase, proses data mining ialah tahap vital dalam prosedur pencarian informasi. Berikut ini adalah tahapan proses pengolahan informasi [21] :

1. Intergrasi Data (Data Intergration) : Tahap pertama semua data di susun dan digabungkan dari berbagai sumber.
2. Pemilihan Data (Data Selection) : Karena semua data yang dikumpulkan oleh pengguna tidak sepenuhnya diperlukan. Pada tahap ini pemilihan data yang dianggap bermanfaat untuk penambangan data.
3. Pembersihan Data (Data Cleaning) : Data yang tersimpan belum tentu bersih. Ini mungkin masih mengandung kesalahan, nilai yang hilang atau data yang belum dipercaya. Jadi pada thap ini digunakan berbagai metode untuk membersihkan dari ketidak konsistenan tersebut.
4. Transformasi Data : Modifikasi data menjadi bentuk yang quired untuk operasi penambangan data disebut Transformasi Data.
5. Data Mining : Terdiri dari berbagai Teknik yang dapat digunakan untuk menemukan jalan keluar dari berbagai pengaturan atau kesamaan yang tidak teramati dalam dataset yang diberikan.
6. Evaluasi pola dan Presentasi pengetahuan (Pattern Evaluation and Knowledge Presentation) : langkah ini termasuk menghilangkan atau menghapus pola duplikat dari pola yang kami hasilkan.
7. Keputusan/penggunaan pengetahuan yang ditemukan (Decisions/Use of Discovered Knowledge) : langkah ini membantu pengguna untuk membuat keputusan tentang informasi yang dikumpulkan.

## Text Mining

Text mining adalah proses menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen dan tujuannya adalah mencari kata – kata yang dapat mewakili isi dokumen sehingga dapat dilakukan analisis keterhubungan antar dokumen tersebut [20].

Text mining juga dikenal sebagai data mining text atau penemuan pengetahuan dari database tekstual. Sesuai dengan buku The Text Mining Handbook, text mining dapat didefinisikan sebagai suatu proses menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan tools analisis yang merupakan komponen – komponendalam data mining. Tujuan dari text mining adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Jadi, sumber data yang digunakan dalam text mining adalah sekumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi terstruktur. Adapun tugas khusus dari text mining antara lain yaitu pengkategorisasian teks dan pengelompokkan teks [22].

## Twitter

Twitter adalah suatu situs web yang merupakan layanan dari microblog, yaitu suatu bentuk blog yang membatasi ukuran setiap posnya, yang memberikan fasilitas bagi pengguna untuk dapat menuliskan pesan dalam twtiiter update hanya berisi 140 karakter. Twitter merupakan salah satu jejaring sosial yang paling mudah digunakan, karena hanya memerlukan waktu yang singkat tetapi informasi yang disampaikan dapat langsung menyebar secara luas [26]. Kemudian fitur fitur yang terdapat dalam Twitter ini sendiri ialah :

1. Halaman Utama (*Home*)

Merupakan halaman utama untuk melihat tweets yang dikirimkan oleh orang-orang yang berteman dengan kita atau yang kita ikuti (Following).

1. Profil (*Profile*)

Halaman ini yang bisa dilhat oleh semua orang yang berisi profl atau data diri serta tweets yang sudah pernah dituliskan atau pernah diupdate.

1. Followers

Disebut juga pengikut, merupakan pengguna lain yang ingin menjadikan teman. Pengguna lain yang menjadi pengikut akun seseorang, maka Tweets seseorang yang ia ikuti tersebut akan muncu di halaman utamanya.

1. Following

Merupakan akun seseorang yang mengikuti akun pengguna lain agar Tweets yang dikirim oleh orang yang diikuti bisa masuk ke dalam halaman utam twitternya.

1. Penyebutan (*Mention*)

Isitilah penyebutan (Mentions) biasanya konten ini merupakan balasan dari percakapan agar sesama pengguna bisa langsung menandai dengan mudah orang yang akan diajak bicara.

1. Favorit (*Favorite*)

Fitur ini biasanay digunakan untuk menyimpan tweets yang disukai agar tidak tenggelam dengan Tweets lain.

1. Pesan Langsung (*Direct Message*)

Bisa dibilang layaknya SMS karena pesan yang dikirim akan langsung masuk ke akun seseorang yang akan dikirim secara personal.

1. Tanda Pagar (*Hashtag*)

Simbol “#” atau bisa dibilang hashtag yang ditulis didepan suatu topik tertentu yang bertujuan agar pengguna lain atau akun lain bisa mencari topik yang sama sesuai yang dicari dengan hashtag itu.

1. Daftar (*List*)

Pengguna twitter dimudahkan untuk dapat melihat secara keseluruhan para nama pengguna (username) yang mereka ikuti (follow).

1. Topik Terkini (Trending Topic)

Hal ini bisa dilihat berdasar rating rang yang membicarakan topik yang sama dalam waktu bersamaan dan didalam daerah itu sendiri.

## Preprocessing

*Preprocessing* merupakan proses untuk mempersiapkan data sebelum dilakukan proses lebih lanjut, data awal yang diperoleh biasanya masih bagian-bagian yang tidak diperlukan pada analisa sentimen, oleh sebab itu perlu dilakukan preprocessing. terutama data yang diperoleh dari media sosial yang banyak mengandung kata-kata yang tidak formal dan tidak terstruktur sehingga menimbulkan noise yang besar, noise yaitu kesalahan acak atau varian dalam variabel terukur [27].

## Seleksi Fitur

Seleksi fitur adalah salah satu teknik terpenting dan sering digunakan dalam preprocessing. Teknik ini mengurangi jumlah fitur yang terlibat dalam menentukan suatu nilai kelas target, mengurangi fitur irelevan, berlebihan dan data yang menyebabkan salah pengertian terhadap kelas target yang membuat efek segera bagi aplikasi. Tujuan utama dari seleksi fitur ialah memilih fitur terbaik dari suatu kumpulan fitur data [24].

Seleksi fitur merupakan salah satu faktor yang paling yang dapat mempengaruhi tingkat akurasi klasifikasi karena jika dataset berisi jumlah fitur, dimensi ruang akan menjadi besar dan merendahkan tingkat akurasi klasifikasi. Seleksi fitur juga sangat mempengaruhi beberapa aspek yaitu pola klasifikasi, akurasi klasifikasi, waktu yang diperlukan untuk pembelajaran fungsi klasifikasi, jumlah contoh yang dibutuhkan untuk pembelajaran dan biaya yang terkait dengan fitur [25].

## Naïve Bayes Classifier

Naive Bayes Classifier adalah salah satu metode klasifikasi atau algoritma klasifikasi berbasis probability kemunculan fitur-fitur datasetnya terhadap fitur yang digunakan untuk training. Naïve Bayes dapat diartikan sebuah metode klasifikasi yang berdasarkan probabilitas, dengan asumsi bahwa setiap variable x bersifat bebas (independent) [31].

Naive Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma mengunakan teorema Bayes dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas [33].

## Adaboost

Algoritma AdaBoost pertama kali diperkenalkan pada tahun 1995 oleh Freund dan Schapire, telah banyak memecahkan berbagai masalah praktis dari algoritma boosting sebelumnya, pada dasarnya metode boosting (Adaboost) dapat meningkatakan ketelitian dalam proses klasifikasi dan prediksi dengan cara membangkitkan kombinasi yang dipilih adalah model yang meiliki nilai bobot paling besar [29].

AdaBoost (Adaptive Boosting) merupakan algoritma machine learning yang dirumuskan oleh Yoav Freund dan Robert Schapire [29] [30]. Algoritama AdaBoost merupakan algoritma yang membangun pengklasifikasi kuat dengan mengombinasikan sejumlah pengklasifikasi sederhana (lemah).

# METODOLOGI

## Sumber dan Tipe Data

Data yang digunakan pada penelitian ini adalah data sekunder yang berupa *review* produk Scarlett Whitening yang terdapat pada Twitter.

## Metode Pengolahan Data

Tahap-tahap yang dilakukan dalam melakukan penelitian ini secara umum adalah :

1. Mengumpulkan Data
2. Menentukan Atribut
3. Melakukan Proses Data Mining
4. Pemilihan Data Training
5. Mencari kelayakan metode Naive Bayes dengan menghitung probabilitas kelas
6. Menghitung probabilitas masing-masing atribut
7. Menghitung probabilitas akhir untuk setiap kelas
8. Melakukan Data Testing

## Diagram Alur Penelitian

# JADWAL PENELITIAN

Penelitian direncanakan akan dilaksanakan selama enam bulan. Rincian rencana jadwal penelitian dicantumkan dalam Tabel 1.

Tabel 1. Jadwal penelitian.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| No. | Keterangan | Bulan | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 |
| 1. | Telaah literatur |  |  |  |  |  |  |
| 2. | Pengadaan bahan |  |  |  |  |  |  |
| 3. | Pembuatan purwarupa |  |  |  |  |  |  |
| 4. | Pengambilan data |  |  |  |  |  |  |
| 5. | Kompilasi dan analisis data |  |  |  |  |  |  |
| 6. | Evaluasi dan perbaikan purwarupa |  |  |  |  |  |  |
| 7. | Penulisan Laporan |  |  |  |  |  |  |