BAB III

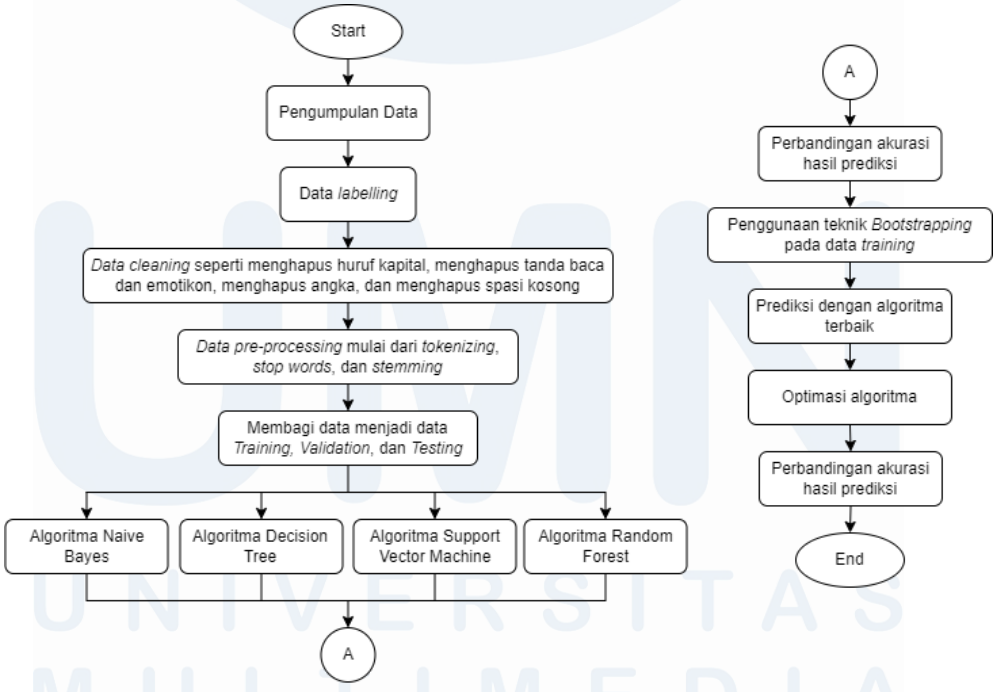
METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Objek penelitian kali ini adalah data ulasan atau *review* dari Google Play Store terhadap aplikasi Bibit untuk menentukan bagaimana perkembangan pendapat para pengguna aplikasi Bibit. Data tersebut kemudian diproses terlebih dahulu sebelum membagi sentimen pengguna ke dalam 2 kategori yaitu positif dan negatif. Berikutnya, analisis sentimen akan dilakukan dengan menggunakan algoritma *NB, DT, SVM,* dan *RF*. Data ulasan yang diambil sebagai data pada penelitian kali ini berada pada rentang waktu tahun 2020 sampai 2022 sehingga dari 3 tahun tersebut dapat dianalisis pula perkembangan ulasan pengguna dari tahun ke tahun khususnya pada saat Covid-19 melanda.

3.2 Metode Penelitian

3.2.1 Alur Penelitian



Gambar 3.1 Diagram alur penelitian

Gambar 3.1 merupakan diagram alur atau *flowchart* yang menjelaskan mengenai apa saja proses yang dilakukan pada penelitian kali ini. Pertama-

tama data ulasan aplikasi Bibit dari Google Play Store diambil terlebih dahulu. Selanjutnya, dilakukan data *labeling* yang akan menjadi patokan dalam melatih algoritma. Setelah itu, dilakukan pembersihan data seperti penghapusan huruf kapital (*case folding*), penghapusan tanda baca dan emotikon (*punctuation and emoji*), penghapusan angka (*numbers*), dan penghapusan spasi kosong (*whitespace*). Proses *data cleaning* dilanjutkan dengan *data pre-processing* seperti *tokenizing*, *stop words*, dan *stemming*. Berikutnya data akan dibagi ke dalam data *training* dan *testing* diikuti dengan *TF-IDF* untuk memberikan informasi kepada model yang akan dilatih nanti mengenai seberapa sering kata tersebut muncul pada ulasan pengguna. Data *training* akan digunakan untuk melatih model sementara data *testing* akan digunakan untuk menguji coba kemampuan model untuk melakukan prediksi dengan tepat.

Tahapan selanjutnya adalah membuat model dengan 4 algoritma yaitu *NB*, *DT*, *SVM*, dan *RF* menggunakan data yang sudah dibagi sebelumnya ke dalam *training* dan *testing*. Keempat algoritma tersebut dilatih terlebih dahulu dengan menggunakan data *training* agar mampu mempelajari pola dan kata apa saja yang termasuk ke dalam sentimen positif dan sentimen negatif. Hasil akurasi akan diukur dengan menggunakan *confusion matrix* dan dipilih model terbaik dari 4 algoritma tersebut. Algoritma yang terpilih akan dilatih ulang dengan menggunakan teknik *bootstrapping*. Umumnya, terdapat perbedaan yang signifikan antar kelas berbeda pada data *training*. Ketidakseimbangan itu dapat menimbulkan bias pada model atau kecenderungan untuk memilih kelas mayoritas pada saat melakukan prediksi. Oleh karena itu, teknik *bootstrapping* dilakukan untuk menyeimbangkan jumlah data *training* pada masing-masing kelas sentimen positif dan negatif dengan cara menduplikasi secara acak kelas data minoritas.

Penelitian dilanjutkan dengan melakukan optimasi untuk melihat apakah terdapat parameter yang lebih baik daripada parameter algoritma standar. Optimasi ini dilakukan dengan menggunakan *GridSearchCV*. Cara penggunaan *GridSearchCV* ini adalah dengan menambahkan batasan atau area pada parameter-parameter yang ingin diubah dari algoritma yang digunakan.

GridSearchCV kemudian akan menjalankan algoritma tersebut dengan kombinasi setiap parameter yang didefinisikan sebelumnya. Setelah itu, kombinasi parameter dengan akurasi terbaik akan disimpan dan dijadikan model baru. Hasil akurasi dengan model baru ini akan dibandingkan lagi dengan akurasi model standar dengan algoritma terpilih sebelumnya. Model terbaik akan digunakan dalam pembuatan *website* sebagai hasil akhir penelitian kali ini.

3.2.2 Metode Pengembangan Sistem / Metode Data Mining

Metode *data mining* yang digunakan pada penelitian kali ini adalah CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Penelitian ini memiliki 6 tahapan yang akan diaplikasikan dengan cara berikut ini:

1. *Business Understanding*

   Tahapan ini bertujuan untuk mengetahui mengapa penelitian ini harus dilakukan, apa manfaatnya, dan masalah apa yang ingin diselesaikan. Dalam hal ini, penelitian ini dilakukan dengan tujuan untuk melihat perkembangan sentimen aplikasi Bibit mulai dari tahun 2020 hingga 2022. Selain itu, hasil analisis sentimen tersebut dapat dijadikan sebagai rekomendasi bagi calon pengguna baru aplikasi Bibit.

2. *Data Understanding*

   Data understanding merupakan tahapan untuk memahami data yang diperoleh secara menyeluruh meliputi proses mengumpulkan data, eksplorasi data, mendeskripsikan data, dan melakukan pengecekan terhadap kualitas data [36]. Penelitian ini menggunakan data dari ulasan atau *review* Google Play Store terhadap aplikasi Bibit. Data diperoleh dengan menggunakan *library google-play-scraper* kemudian data tersebut difilter dengan rentang waktu yang dipilih, yaitu dari tanggal 1 Januari 2020 sampai dengan 31 Desember 2022. Terdapat beberapa kolom data ulasan yang diambil dari Google Play Store namun, mengingat penelitian ini hanya akan melakukan analisis sentimen maka hanya kolom *content* berisi ulasan-ulasan yang akan digunakan dan *score* yaitu *rating* pengguna

terhadap aplikasi Bibit yang akan digunakan untuk pemberian label sentimen positif dan negatif.

3. *Data Preparation*

   Data preparation merupakan tahapan untuk mempersiapkan data terlebih dahulu agar siap diproses di tahapan selanjutnya. Langkah yang diambil untuk mempersiapkan data untuk penelitian ini adalah dengan melakukan *data cleaning* dengan menggunakan Python seperti mengubah huruf kapital menjadi huruf kecil atau *case folding*, menghapus emotikon dan tanda baca, menghapus angka, dan menghapus spasi kosong atau *whitespace*. Selanjutnya, data diberikan label sesuai dengan *rating* 1-5 dari pengguna. *Rating* 1 dan 2 akan diberikan label sentimen negatif sementara itu *rating* 4 dan 5 akan diberikan label sentimen positif. Ulasan dengan *rating* 3 akan dihapus karena jumlah datanya terlalu sedikit sekaligus untuk menghindari terjadinya ambiguitas pada saat melatih model dengan algoritma-algoritma yang digunakan. Proses *data preparation* dilanjutkan dengan melakukan *data pre-processing*. Tahapan *data pre-processing* yang dilakukan meliputi *tokenizing, stop words*, dan *stemming*. *Tokenizing* dilakukan dengan tujuan mengubah sebuah kalimat penuh menjadi kata-kata. *Stop words* akan menghapus kata yang tidak memiliki pengaruh dalam kalimat sentimen seperti misalnya kata penghubung. Terakhir, proses *stemming* bertujuan untuk mengubah kata-kata yang ada menjadi kata dasar bakunya. Hasil akhir setelah pembersihan data ulasan yang tersisa berjumlah 34.230 dari total 36.670 ulasan selama tahun 2020 sampai 2022.

4. *Data Modeling*

   Tahapan *data modeling* meliputi proses pemilihan model serta pembangunan model sampai model terbentuk [36]. Pertama-tama akan dilakukan pemisahan data utama menjadi data *training* dan *testing* yang diikuti dengan *feature extraction* menggunakan *TF-IDF* (*Term Frequency-Inverse Document Frequency*) yang berfungsi untuk memberikan bobot frekuensi setiap kata yang akan dianalisis. Selanjutnya, pembuatan model akan dilakukan dengan 4 algoritma yaitu *NB*, *DT*, *SVM*, dan *RF*.

5. *Evaluation*

   Tahapan *evaluation* berfungsi untuk mengevaluasi hasil atau performa model-model yang sudah terbentuk dari tahapan *data modeling*. Proses evaluasi ini akan menggunakan *confusion matrix* sehingga hasil akurasi dari keempat model dapat terlihat dengan mempertimbangkan *precision*, *recall*, *f1-score*, dan *accuracy*. Berdasarkan 4 model sebelumnya, dipilih salah satu model yang terbaik untuk dioptimasi lebih lanjut menggunakan teknik *bootstrapping* dan *GridSearchCV*. Hasil akhirnya akan dikomparasi ulang dan dipilih model dengan parameter terbaik untuk dilanjutkan ke tahapan *deployment*.

6. *Deployment*

   Deployment merupakan tahapan terakhir dari metode CRISP-DM. Pada tahapan ini, model yang sudah dibentuk sebelumnya akan di *deploy* pada sebuah *website* sederhana. Melalui *website* tersebut, pengguna dapat mencoba menggunakan model yang dipilih untuk memprediksi sentimen seperti ulasan untuk mengetahui apakah ulasan tersebut masuk ke dalam kategori positif atau negatif.

3.3 Teknik Pengumpulan Data

3.3.1 Populasi dan Sampel

Populasi yang akan diambil untuk penelitian kali ini adalah ulasan dari Google Play Store pada aplikasi Bibit. Teknik sampling yang digunakan adalah teknik *non-probability* sampling yaitu *purposive sampling*. *Purposive sampling* merupakan teknik sampling yang dilakukan dengan menentukan terlebih dahulu tujuan dari penelitian yang dilakukan. Oleh karena itu, *purposive sampling* memiliki keunggulan dalam hal data yang diambil lebih akurat dan hasil penelitian juga menjadi lebih optimal jika dibandingkan dengan teknik pengambilan sampel yang lain [43]. Sampel diambil dengan menggunakan *google-play-scraper* melalui Google Colaboratory yang memungkinkan penggunaan dengan bahasa Python.

3.3.2 Periode Pengambilan Data

Pada penelitian kali ini periode pengambilan data ulasan aplikasi Bibit di Google Play Store berada pada rentang waktu 1 Januari 2020 sampai 31 Desember 2022. Alasan dipilihnya rentang waktu 2020 sampai 2022 karena selama 3 tahun tersebut data ulasan dapat diperoleh selama 12 bulan penuh sehingga jumlah data setiap tahunnya diharapkan lebih seimbang. Oleh karena itu, nantinya dapat dibandingkan perkembangan sentimen pengguna aplikasi Bibit dari tahun ke tahun. Total terdapat 36.670 ulasan yang diambil dari Google Play Store.

3.4 Variabel Penelitian

Dalam penelitian kali ini, terdapat dua buah variabel penelitian yaitu variabel dependen dan variabel independen.

1. Variabel Dependen

   Variabel dependen adalah variabel yang memiliki keterikatan atau perubahannya bergantung kepada variabel independen. Pada penelitian ini, variabel dependennya adalah hasil label sentimen positif dan negatif.

2. Variabel Independen

   Variabel independen merupakan variabel yang menyebabkan adanya perubahan pada variabel dependen. Berdasarkan data ulasan yang dimiliki maka variabel independen adalah kumpulan komentar atau ulasan mengenai aplikasi Bibit yang diambil dari Google Play Store.

3.5 Teknik Analisis Data

Data pada penelitian ini bersifat kualitatif karena berbentuk ulasan sehingga penelitian akan dilakukan secara kualitatif. Analisis dilakukan dengan menggunakan Python sebagai bahasa pemrogramannya dengan tools Google Colaboratory yang memungkinkan pengerjaan kapan pun dan di mana pun karena terkoneksi dengan akun Google. Data akan melalui tahapan persiapan terlebih dahulu atau disebut juga *pre-processing*, dilanjutkan dengan *feature extraction* menggunakan *TF-IDF* (*term frequency-inverse document frequency*) dan analisis menggunakan 4 algoritma yaitu *NB, DT, SVM,* dan *SVM.* Setelah itu, penelitian

akan ditutup dengan pengujian akurasi performa algoritma menggunakan *confusion matrix*. Terdapat beberapa visualisasi data yang nantinya menggunakan *tools* Microsoft Excel.