BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Objek penelitian yang diteliti adalah pengguna Quora yang telah menuliskan ulasan mereka di tempat ulasan Google Play Store dengan khusus untuk *region* Indonesia. Quora merupakan aplikasi yang digunakan untuk menulis pertanyaan dan menjawab pertanyaan yang diberikan oleh pengguna Quora lainnya. Ulasan Quora tersebut diteliti kemudian hasil tersebut akan menghasilkan sentimen pengguna aplikasi Quora di Indonesia berupa positif atau negatif. *Dataset* berbentuk *csv* dikumpulkan melalui metode *scrapping* pada *website* Google Play Store Id yang berisi 1500 data dengan prioritas *most relevant*.

3.2 Metode Penelitian

Untuk metode penelitian ini, tipe algoritma yang digunakan adalah *supervised learning* dengan metode klasifikasi yaitu *decision tree* dan *K-Nearest Neighbors*. *Decision Tree* dipakai karena *decision tree* dapat memberikan intepretasi atas pikiran masyarakat yang lebih baik karena logika dari *decision tree* mirip dengan manusia. Selain itu *decision tree* dan *K-Nearest Neighbors* cocok untuk visualisasi data dan mudah digunakan. *Decision tree* juga dapat digunakan pada *data* numerik dan kategori karena hasil dari sampel dan dataset tidak selalu memberikan data numerik terutama yang berhubungan dengan hal-hal subjektif seperti minat seseorang. *Decision tree* dan *K-Nearest Neighbors* menghasilkan keputusan yang sangat penting untuk evaluasi bagi pengembang Quora. Penggunaan *K-Nearest Neighbors* disini juga dapat melakukan analisis sentimenn yang mempunyai dua kelas atau lebih seperti halnya sentimen analisis yang mempunyai 2 sentimen yaitu positif dan negatif. Dataset yang telah didapat akan dilakukan *data training* dan *data testing* menggunakan aplikasi pengolah data. Setelah hasil dari data tersebut didapatkan maka kedua algoritma akan dibandingkan dan hasil data tersebut akan divisualisasikan agar hasil analisis data yang telah diolah dapat

ditampilkan serta dipahami oleh orang awam. Visualisasi data statistik dari dataset akan ditampilkan dalam berbagai *chart* atau *model*.

3.3 Teknik Pengumpulan Data

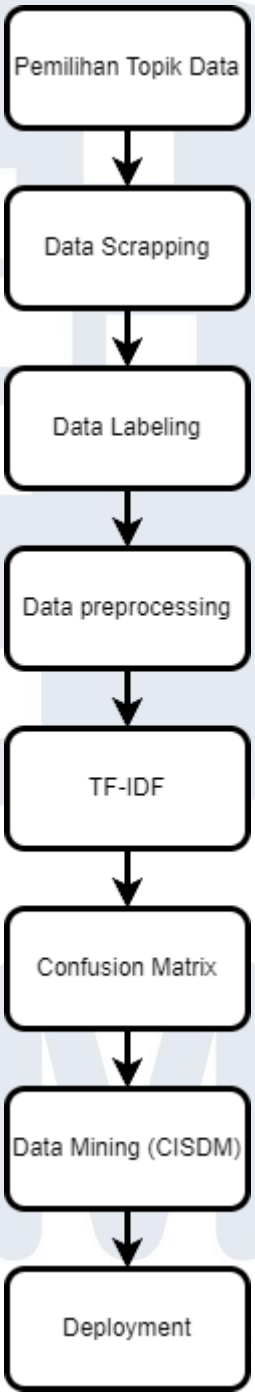
Untuk sumber pengumpulan data akan mengandalkan teknik pengumpulan data dengan sumber langsung yaitu sumber primer. Teknik yang digunakan adalah dengan menggunakan dataset. *Dataset* yang diambil merupakan dataset dari berbagai *review* pengguna Quora *region* Indonesia. Jenis data yang akan digunakan adalah data kuantitatif. Metode kuantitatif merupakan metode yang mengandalkan kuantitas atau jumlah dari responden untuk mendapatkan data. Terdapat 1500 data review yang diambil menggunakan *data scrapper* dari Google. Proses *scrapping* memakan waktu sekitar 10 detik - 15 detik. Proses pengambilan data dilakukan pada bulan 18 Januari 2024 dikarenakan jumlah data yang bisa diambil terbatas ketika proses scrapping dilakukan berulang kali sehingga pengambilan data dilakukan pada masa awal-awal proses pengerjaan skripsi untuk mengambil penuh 1500 dataset. Data review Quora tersebut diambil dari 2 tahun kebelakang yaitu dari tahun 8 April 2022 sampai 16 Januari 2024.

3.4 Teknik Pengambilan Sampel

Untuk Teknik pengambilan sampel digunakan teknik pengambilan sampel acak (*probability*). Tipe sampel *probability* yang diambil adalah random sampling yaitu data akan diambil secara acak tanpa adanya karakteristik secara khusus.

3.5 Alur Penelitian

Alur penelitian merupakan langkah-langkah atau tahapan yang harus dilalui oleh peneliti untuk melakukan penelitian berdasarkan langkah-langkah yang telah ditentukan berdasarkan dengan standar yang berlaku sehingga penelitian yang dilakukan menjadi *valid* dan konkrit. Alur penelitian ini terbagi menjadi beberapa langkah mulai dari proses penentuan topik *data*, pengumpulan *data* atau yang biasa disebut dengan *scrapping* sampai dengan proses *deployment*. Berikut ini adalah diagram alur penelitian sentimen analisis ini [18].



Gambar 3. 1 Diagram Alur Penelitian

3.5.1 Pemilihan Topik *Data*

Tahap pertama ini merupakan pemilihan topik *data*. Tahap ini merupakan proses pemilihan data yang sesuai dengan jenis penelitian yang dilakukan saat ini. Karena jenis penelitian ini merupakan sentimen analisis, maka topik *data* yang dicari adalah *data* yang berhubungan dengan opini. Banyak opini yang dapat dipilih untuk keperluan analisis sentimen ini seperti opini pada sosial media ataupun *review* suatu benda tertentu. Setelah mencari topik dan melakukan pertimbangan, maka diputuskan bahwa topik *data* tersebut mengenai analisis sentimen terhadap aplikasi Quora di Google Play Store.

Topik data yang telah ditentukan memerlukan algoritma yang sesuai agar proses penelitian ini mendapatkan hasil yang memuaskan sesuai dengan alur penelitian yang telah ditentukan. Setelah melalui proses pertimbangan, maka algoritma yang telah diputuskan adalah algoritma *Decision Tree* dan *K-Nearest Neighbors.*

3.5.2 Data Scrapping

Setelah topik data telah ditentukan, langkah selanjutnya adalah *proses scrapping data*. Proses scrapping data merupakan proses pengambilan data dari suatu *website* menggunakan *extension* atau *tools* tertentu. Untuk proses *scrapping* tersebut diputuskan untuk menggunakan *library* pada Jupyter Notebook dengan nama Google Play Scrapper. Tools tersebut mampu mengambil data hasil *review* di Google Play Store. Selain itu, library ini memiliki fitur untuk menentukan jumlah data yang ingin diambil, bahasa, *region*, dan tipe *review* (*most relevant* atau *newest review*). Data yang telah melalui tahap *scrapping* tersebut akan disimpan kedalam bentuk *csv* agar data tersebut dapat dipakai pada *notebook* yang baru atau bisa dipakai dikemudian hari.

3.5.3 Data Labeling

Data yang sudah diunduh daam bentuk *csv* kemudian akan melalui tahapan berikutnya yang dinamakan sebagai data labeling. Data *labeling* ini merupakan proses indentifikasi data yang masih belum diolah menjadi penanda suatu informasi yang dapat dijadikan sebagai konteks bagi data itu sendiri.

Pada tahapan penelitian untuk melakukan sentimen analisis ini, tipe data dapat dibagi menjadi 2 yaitu positif dan negatif. Untuk data mengenai *review* Quora ini maka *column* yang terkena proses *labeling* adalah *score* atau *rating*. Jika rating tersebut lebih dari tiga (>3) maka data tersebut akan dikenakan *labeling* positif. Jika rating tersebut kurang dari tiga (<3) maka data tersebut akan dikenakan *labeling* negatif.

3.5.4 Data Preprocessing

Data yang sudah melalui tahapan *sentiment labeling* akan memasuki tahap berikutnya yaitu tahap *data preprocessing*. Pada tahapan ini, *data* akan diseleksi dan diolah lebih lanjut sehingga *data* tersebut dapat dengan layak dipakai untuk proses *data mining*. Terdapat beberapa langkah yang harus dilakukan dalam melakukan data preprocessing ini.

Langkah pertama adalah dengan menghapus elemen-elemen tertentu pada kalimat *review* yang tidak diperlukan untuk proses *data mining*. Elemen-elemen yang tidak diperlukan tersebut seperti huruf kapital, *URL, mention*, tanda baca, area kosong ditengah kalimat, dan juga *hashtag*.

Langkah kedua setelah penghilangan elemen kalimat adalah proses *tokenizing*. Pada tahapan ini, kalimat *review* yang telah diolah pada langkah sebelumnya akan dipisahkan per katanya menggunakan koma sehingga setiap kata tersebut dapat menjadi data yang dapat di proses.

Setelah melalui tahap *tokenizing*, maka data akan masuk kedalam tahap *filtering*. Pada tahap ini, *review* yang sudah dipisahkan menjadi kata perkata akan diseleksi dengan membuang kata-kata yang tidak penting menggunakan

fitur *stopwords*. Kata-kata yang tidak penting dalam Bahasa Indonesia seperti kata penghubung (cth: yang, di).

Semua *review* yang telah melewati tahap *tokenizing* dan *filtering* kemudian akan memasukin tahapan yang bernama *stemming*. *Stemming* merupakan proses penyederhanaan kalimat atau kata dengan mengembalikan kalimat kedalam bentuk kata dasar. Dalam hal ini misalkan adalah kata imbuhan (cth: pe-an, meng-an) yang diubah menjadi kata kerja dasar tanpa inbuhan. Untuk melakukan *stemming* ini membutuhkan *library* yang bernama "Sastrawi" yang dapat melakukan *stemming* untuk kalimat bahasa Indonesia.

3.5.5 TF-IDF

Setelah melewati tahapan data *preprocessing*, data sudah siap di olah dan digunakan. Data tersebut akan memasuki tahap selanjutnya yaitu TF-IDF. TF-IDF merupakan singkatan dari *Term Frequency* (TF) dan *Document Frequency* (DF). TF-IDF ini merupakan metode yang mengubah kata atau kalimat menjadi sebuah bobot angka yang dapat dibaca oleh mesin sehingga dapat dijalankan oleh algoritma klasifikasi maupun prediksi. TF-IDF ini dapat digunakan untuk meningkatkan akurasi model klasifikasi dan menemukan kata kunci dokumen sehingga kata yang paling relevan dapat diidentifikasi.

3.5.6 Confusion Matrix

Sebelum *data mining* dilakukan maka dilakukanlah *confusion matrix*. *Confusion matrix* ini digunakan untuk mengukur kinerja algoritma klasifikasi. Terdapat 4 hasil yaitu

1. *True Positive* (TP): Jumlah data yang Positif dan benar diprediksi sebagai Positif.
2. *False Positive* (FP): Jumlah data yang bernilai Negatif tetapi diprediksi sebagai Positif.

3. *False Negative* (FN): Jumlah data yang bernilai Positif tetapi diprediksi sebagai Negatif.
4. *True Negative* (TN): Jumlah data yang bernilai Negatif dan benar diprediksi sebagai Negatif.

3.5.7 Data Mining

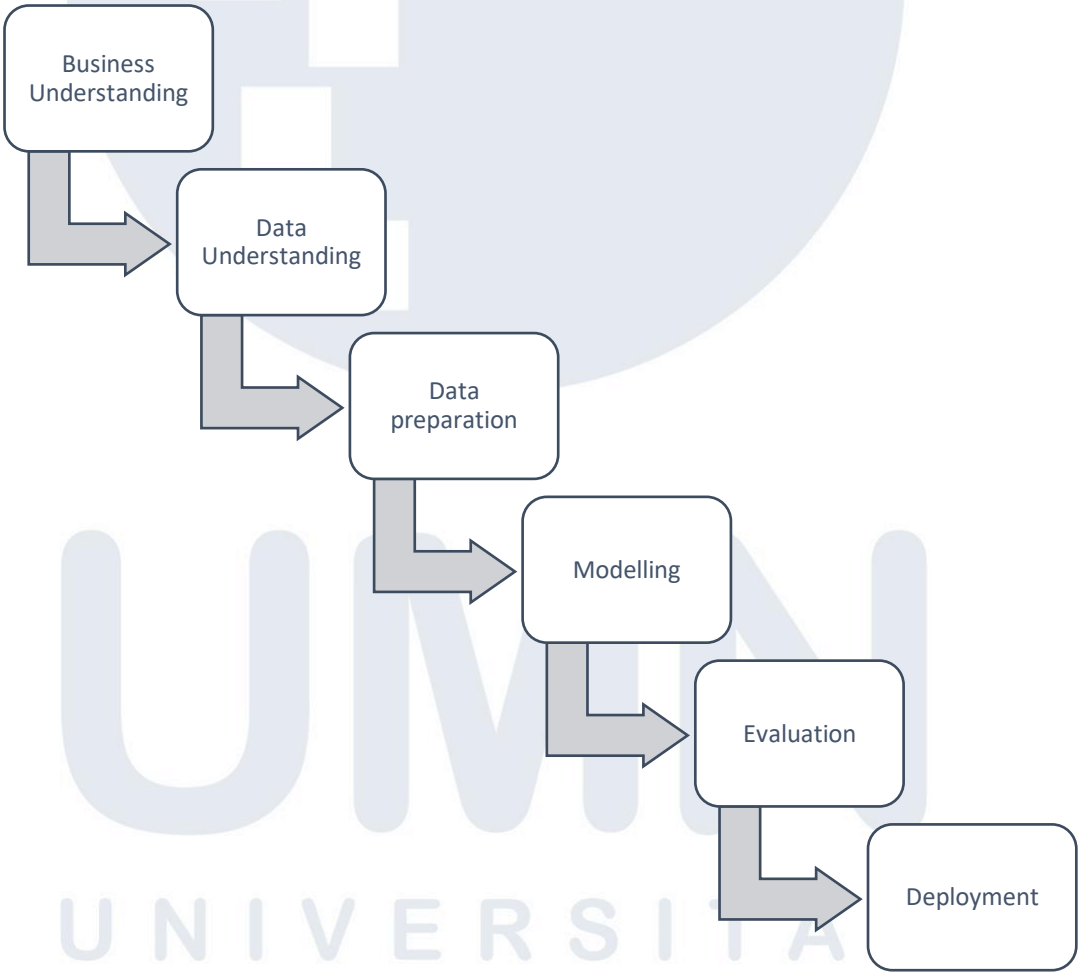
Pada tahap ini, data sudah benar-benar siap digunakan untuk menjalankan *data mining* menggunakan algoritma yang ditentukan yaitu K-Nearest Neighbors dan Decision Tree. Metode *data mining* ini menggunakan teknik CISDM (*Cross Industry Standard Process for Data Mining)* yang diawali dengan pemahaman akan data, persiapan data, pemrosesan data, *modelling* untuk pencarian akurasi sampai proses *deployment* untuk menyajikan data. Sebelum melakukan pengujian algoritma, maka data akan melalui tahap *splitting* yaitu membagi data kedalam *training* dan *test* terlebih dahulu untuk menguji algortima. Hasil akhir algoritma akan menghasilkan akurasi untuk *precision, recall, f1-score,* dan *support.* Hasil akhir tersebut kemudian akan dibandingkan dengan algoritma yang lain untuk menemukan hasil yang terbaik.

3.5.8 Deployment

Setelah data menghasilkan akurasi dan *confusion matrix*, maka data akan memasuki tahap deployment yaitu hasil akhir dari algoritma akan ditampilkan dalam bentuk plot sehingga lebih mudah dipahami oleh orang awam seperti *heat map* dan *barplot.* Selain itu, hasil analisis sentimen akan ditampilkan dalam bentuk *wordcloud*. *Wordcloud* merupakan tampilan banyak kata-kata yang menampilkan kata paling positif dan negatif dari yang terbesar hingga yang paling kecil. Data juga akan dipresentasikan kedalam *website* yang mampu memudahkan pengguna dalam mencari dan membaca hasil dari analisis sentimen yang telah dilakukan.

3.6 Teknik *Data Mining*

Untuk penelitian ini, teknik yang digunakan untuk menganalisis data secara kuantitatif di mana data tersebut dianalisis dari segi jumlah atau kuantitas data. Metode yang dilakukan untuk melakukan proses alur *data mining* dengan menggunakan CRISP-DM. CRISP-DM merupakan singkatan dari *Cross Industry Standard Process for Data Mining*. CRISP-DM berbentuk seperti alur atau siklus yang akan memproses data dari tahapan paling awal di mana data tersebut masih belum diolah dan belum terstruktur hingga menjadi hasil nyata yang dapat diimplementasikan.



Gambar 3. 2 Diagram Alur CISDM

3.6.1 Business Understanding

Business understanding merupakan proses untuk melihat keseluruhan dari objek dan target yang ingin dianalisa. *Business Understanding* diperlukan untuk mempersiapkan teknik analisa sehingga dapat menyiapkan skenario dan langkah yang tepat dalam penelitian. Topik utama penelitian ini sentimen analisis terhadap review Google Play Store *region* Indonesia mengenai aplikasi Quora.

3.6.2 Data Understanding

Pada tahap data understanding ini, data akan dikumpulkan melalui proses yang bernama *scrapping*. *Dataset* yang sudah mengalami tahap *scrapping* akan dianalisa struktur datanya serta akan disesuaikan dengan algoritma yang ada agar proses CRISP-DM ini dapat berjalan secara lancar. Pada proses ini juga diperlukan penemuan *software* yang tepat untuk menyelesaikan analisa ini yaitu *Jupyter Notebook*.

3.6.3 Data Preparation

Tahap ini merupakan tahap untuk mempersiapkan data sehingga *data* dapat dengan layak diolah lebih lanjut. *Dataset* yang dikumpulkan akan berbentuk csv. Terdapat beberapa tahap pada *data preparation* seperti *data preprocessing* untuk memberikan label sentimen pada dataset, menghilangkan elemen *data* yang tidak dibutuhkan untuk analisis sentimen menggunakan *stopwords* dan *filtering*, *tokenizing* untuk membuang kata yang tidak penting, dan juga proses stemming untuk memisahkan kata yang memiliki imbuhan menjadi kata dasar.

3.6.4 Modelling

Modelling merupakan tahapan yaitu data yang sudah dirapikan akan diolah menggunakan algoritma. *Data* yang dirapikan akan melalui tahap *training* terlebih dahulu untuk melatih algoritma sehingga *data* dapat berkembang dan data *testing* yang bertujuan untuk menguji data yang sudah melalui tahap *training* sehingga dapat ditemukan hasil data training yang dapat mencapai efektivitas terbaik untuk menjalankan algoritma. Setelah itu, *data* yang sudah lolos dalam pelatihan dan pengujian akan dianalisa menggunakan algoritma

supervised learning. Algortima yang terpilih adalah *Decision Tree* dan *K-Nearest Neighbors.*

3.6.5 Evaluation

Pada tahap ini, sudah diketahui hasil akhir dari *modelling* terhadap data yang sudah dikumpulkan dengan menggunakan *software* pengolah data. Hasil akhir dari *modelling* ini akan dievaluasi apakah hasil akhir dari penelitian ini dapat memenuhi tujuan penelitian yang telah ditentukan sebelumnya.

3.6.6 Deployment

Tahap ini merupakan tahap di mana hasil akhir sudah siap disajikan dan digunakan untuk kepentingan yang lebih lanjut. Hasil akhir dari *modelling* ini akan divisualisasikan dengan *model plot* dari *Jupyter Notebook* sehingga lebih mudah dipahami oleh khalayak umum.