**BAB III**

**METODOLOGI PENELITIAN**

# **3.1 Tahapan Penelitian**

Langkah awal ketika akan melakukan penelitian yaitu analisis sistem, tujuannya yaitu untuk mengidentifikasi masalah yang ada dan dapat membuat sistem yang diperlukan dengan baik. Permasalahan dalam penelitian ini adalah bagaimana membangun sebuah sistem yang dapat melakukan klasifikasi analisis sentimen komentar atau ulasan dari pengguna aplikasi *Home Credit* menggunakan metode SVM dan K-NN secara otomatis, akurat, dan cepat. Selanjutnya penulis akan mengidentifikasi dan menjelaskan penyebab masalah yang terjadi untuk kemudian diselesaikan.

# **3.2 Instrumen Penelitian**

Instrumen penelitian adalah salah satu proses pengumpulan data yang di dalamnya terfokus ke dalam permasalahan penelitian yang akan diselesaikan, yaitu bagaimana peneliti akan memfokuskan cara pengumpulan data.

Alat atau perangkat yang digunakan untuk pengumpulan data dalam penelitian ini adalah *web scraping* menggunakan pemrograman bahasa *Python* yang kemudian mengambil data dari *Google Play Store*. Hasil dari *scraping* disebut *dataset*, *dataset* yang didapatkan dari *scraping* ini masih berupa *dataset* mentah atau kotor, karena di dalamnya masih terdapat atribut yang tidak akan digunakan, maka *dataset* ini masih harus dilakukan proses pengolahan lagi agar siap digunakan. *Dataset* ini terdiri dari 12 Atribut bawaan HTML yang setiap atribut memiliki penampungan datanya tersendiri.

18

*Dataset* ini mencakup data sentimen atau ulasan para pengguna aplikasi *Home Credit* yang berisikan ulasan komentar dan ulasan negatif.

# **3.3 Metode Pengumpulan Data, Populasi, dan Sampel Penelitian**

## 3.3.1. Metode Pengumpulan Data

Pengumpulan data adalah tahapan awal dalam metodologi penelitian. Sebelum memulai penelitian ditentukan terlebih dahulu jenis data apa yang akan digunakan saat penelitian berlangsung, atau bahan apa saja yang dibutuhkan saat penelitian, dan sumber data yang digunakan untuk penelitian harus dari sumber yang jelas dan akurat. Adapun metode pengumpulan data yang digunakan oleh penulis pada penelitian ini diperoleh melalui proses Web *scraping* dari ulasan *review* *di Google Play Store*. “*Web scraping* adalah proses ekstraksi data dari *website* atau situs internet dan mengubahnya ke dalam bentuk *dataset* yang sudah terstruktur” (Dogucu & Çetinkaya-Rundel, 2020).

## 3.3.2. Populasi

Data yang digunakan untuk penelitian ini merupakan data sekunder, yaitu data yang berupa sentimen atau ulasan *review* berbahasa Indonesia yang berasal dari para pengguna aplikasi *Home Credit* melalui situs *Google Play Store*. Jumlah populasi data yang digunakan untuk penelitian ini adalah **2.845** **ulasan pengguna** aplikasi *Home Credit*, *dataset* diambil dari tanggal **28 Maret 2023 sampai tanggal 24 Mei 2023,** berisikan *score* dan *content* positif dan negatif.

## 3.3.3. Sampel penelitian

Banyaknya jumlah data populasi ini mengharuskan data-data tersebut harus dibagi ke dalam dua kategori, yaitu data latih dan data uji. Data latih adalah data yang digunakan untuk melatih algoritma dalam mencari model pengerjaan yang sesuai, dan data uji adalah data yang akan digunakan untuk menguji model dan mengetahui performa model yang didapatkan setelah tahapan testing. “Data akan di *split* menjadi 80:20 untuk proses latih 80% dan proses uji berjumlah 20% dari jumlah keseluruhan data populasi, tujuan pembagian data ini agar sistem yang digunakan untuk menganalisis sentimen dapat melakukan prosesnya dengan baik” (Muhammadin & Sobari,

2021).

# **3.4 Metode Analisis Data**

Metode analisa data yang di gunakan penulis dalam analisis data ini adalah metode kuantitatif, dan data yang digunakan untuk penelitian ini yaitu *dataset* berupa teks. *Dataset* ini masih mentah, sehingga masih perlu diolah lagi (*preprocessing*) agar sesuai dengan proses klasifikasi analisis sentimen. Metode analisa data yang diajukan penulis untuk menentukan nilai akurasi dari analisis sentimen ini terdiri dari beberapa proses. Tahapan alur penelitian yang dilakukan oleh penulis adalah sebagai berikut:



Sumber: Data penelitian

Gambar III. 1

Alur Penelitian

## 3.4.1. Pengumpulan Data

Pengumpulan *dataset* yang diambil dari ulasan pengguna berbahasa Indonesia dari pengguna aplikasi *Home Credit* pada situs *Google Play Store* dengan teknik web *scraping*. “Web *scraping* yaitu acuan pada proses pengumpulan data atau informasi dari sebuah situs web dengan mengekstrak kontennya. Teknik ini digunakan ketika data yang diinginkan tidak tersedia melalui sumber RSS atau API” (Sembiring & Sari,

2019).

## 3.4.2. *Preprocessing* Data

 Dilakukannya *preprocessing* yaitu untuk menghindari kurangnya kesesuaian data, gangguan data, dan ketidaksesuaian data yang masih kotor. Langkah-langkah *preprocessing* yang digunakan dalam penelitian ini yaitu, *case folding, cleaning, normalize, stopword removal, stemming,*

# Case Folding

*Case folding* adalah proses semua huruf akan diubah menjadi *lowercase* atau huruf

kecil.

# Cleaning

Tahap setelah *Case Folding* yaitu *cleaning*, pada tahap ini dilakukan pembersihan *dataset* dari karakter selain alphabet a-z, seperti emoji, url, karakter lainnya, menghapus spasi berlebih.

# Normalize

Setelah kumpulan kata melewati *cleaning*, selanjutnya sekumpulan kata ini akan melewati *Normalization* atau pengoreksian kata, yaitu proses mengoreksi kata-kata singkatan dan kata baku yang kurang sesuai menjadi kata sesuai, seperti sll jadi selalu, prn menjadi pernah, dan lainnya.

# Stopword Removal

Setiap kata akan diperiksa, jika terdapat kata yang tidak ada hubungannya dengan analisis sentimen atau kata yang tidak mengandung unsur sentimen, maka kata tersebut akan dihilangkan.

# Stemming

*Stemming* merupakan proses untuk menghilangkan atau memotong afiks (akhiran atau awalan), atau menghilangkan imbuhan dari sebuah kata untuk mendapatkan bentuk dasar kata yang lebih sederhana.

## 3.4.3. Data *Split*

“Data *split* fungsinya untuk membagi data menjadi dua bagian, yaitu bagian data latih dan bagian data uji. Data latih digunakan untuk mengetes klasifikator, dan data uji digunakan untuk memperkirakan *error* dari klasifikator yang telah dilatih” (Annur et al., 2022). Ilustrasi pembagian data bisa dilihat di **Gambar III.3**

Total data yang digunakan

|  |  |
| --- | --- |
| Data Latih  | Data Uji  |

Sumber: Data olahan

Gambar III. 2

Pembagian Data latih dan data uji

## 3.4.4. Pembobotan data *Term* TF-IDF

“Pembobotan kata adalah proses untuk mengubah nilai dari kata, frase, atau indeks lainnya(term) lalu memberikan indikator yang sesuai dengan tingkat kepentingannya” (Diandra Audiansyah et al., 2022). Setiap kata (*term*) dari seluruh dokumen yang akan digunakan untuk klasifikasi harus dihitung terlebih dahulu bobot dari setiap kata yaitu menggunakan algoritma TF-IDF *(term frequency, inverse document frequency)*. TF-IDF mempunyai perhitungannya masing-masing, TF (*term frequency*) menghitung kemunculan kata pada setiap dokumen yang terdapat pada suatu kalimat. IDF *(inverse document frequency*) menghitung nilai TF yang didapatkan menggunakan persamaan/rumus yang ada. Berikut persamaan TF-IDF:

 

Sumber: (Aulia et al., 2023)

Gambar III. 3 Persamaan TF-IDF Keterangan:

𝑊𝑡,𝑑 = Nilai TF-IDF

𝑡𝑓𝑡,𝑑 = Nilai kemunculan kata (t) pada dokumen (d)

𝑖𝑑𝑓𝑡 = Nilai *inverse* dokumen setiap kata (t)

𝑑𝑓𝑡 = Nilai kemunculan dokumen setiap kata (t)

N = Nilai seluruh dokumen

## 3.4.5. Implementasi Algoritma

 Setelah melakukan *preprocessing* dan pembobotan kata menggunakan TFIDF, data seharusnya sudah bisa digunakan untuk perhitungan algoritma *Support Vector Machine* dan *K-Nearest Neighbor*. “K-Fold adalah suatu metode validasi silang (*cross-validation*) yang digunakan dalam pengembangan model mesin pembelajaran

(*machine learning*) untuk mengukur performa model secara lebih akurat dan mencegah *overfitting*. K-Fold Cross-Validation melibatkan pembagian data menjadi k subset yang lebih kecil (biasanya k = 5 atau 10)” (Chohan et al., 2020). Semua pengerjaannya dilakukan dengan otomasisasi program dengan bahasa *python*.

## 3.4.6. Evaluasi

 Pengujian penelitian ini untuk mengevaluasi hasil dari pengerjaan algoritma yang digunakan (SVM dan KNN). Metode evaluasi yang digunakan *Confusion Matrix*.

Metode ini membantu untuk memproses hasil analisis *classifier*. Setelah *Confusion Matrix* maka hasil akhirnya bisa dilihat berapa hasil nilai *accuracy, recall, precission*, dan f1-*score*. Perhitungan *Confussion matrix* dilakukan dengan persamaan sebagai berikut: (Aulia et al., 2023).

1. *Accuracy* adalah nilai dari data uji kelas yang diprediksi akurat oleh model dan selaras dengan kelas asli. Akurasi menghitung persentase dari jumlah prediksi yang benar (*true positive* dan *true negative*) dari total jumlah data. Persamaan untuk menghitung akurasi adalah: 
2. Presisi adalah hasil prediksi algoritma antara rasio kesamaan kelas data prediksi algoritma dengan kelas data aslinya. Persamaan untuk menghitung presisi adalah: 
3. *Recall* adalah hasil nilai rasio dari tepatnya model algoritma dalam memprediksi data kelas aslinya adalah: 
4. F-*measure* atau f1-measure adalah perhitungan gabungan antara presisi dan *recall*, persamaan untuk menghitung F-*Measure* adalah: 

Untuk lebih mudah dalam membaca hasil klasifikasi positif dan negatif. *Confusion*

*Matrix* dapat divisualisasikan dalam bentuk tabel.

Tabel III. 1

Tabel *Confussion* *Matrix*



Sumber: (Putri et al., 2023) Dengan keterangan :

1. *True Positive* (TP), adalah jumlah data yang diklasifikasikan positif dan diprediksi dengan benar.
2. *True Negative* (TN), adalah jumlah data yang diklasifikasikan negatif dan diprediksi dengan benar.
3. *False Positive* (FP), adalah jumlah data yang diklasifikasikan negatif tetapi diprediksi sebagai positif.
4. *False Negative* (FN), adalah jumlah data yang diklasifikasikan positif tetapi diprediksi sebagai negatif.