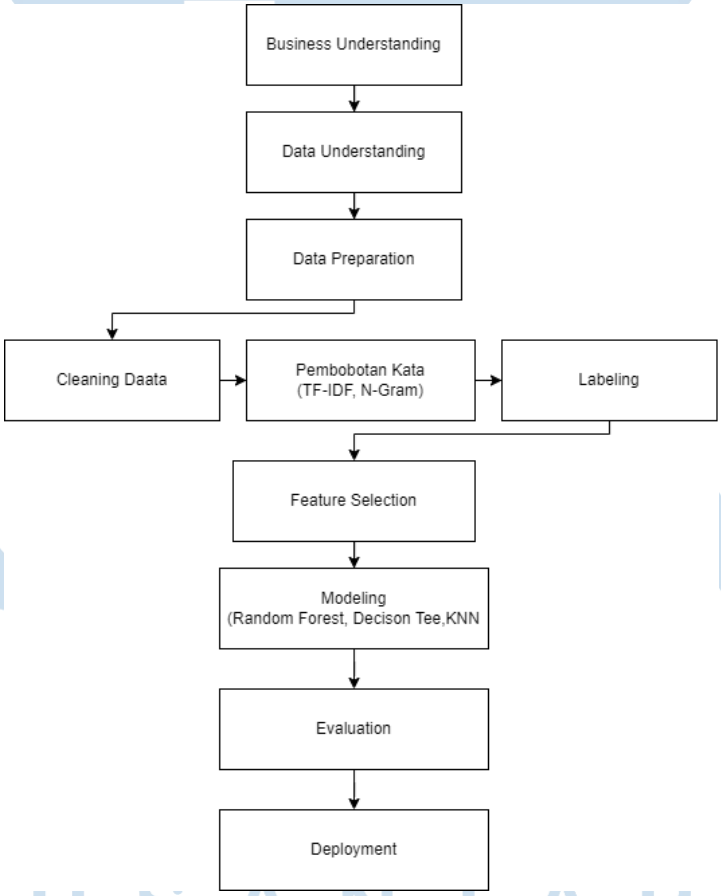
BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Penelitian ini memfokuskan pada objek penelitian untuk melakukan Analisis sentimen pada aplikasi IDN berdasarkan ulasan pengguna *Google Play Store* IDN merupakan aplikasi berita yang paling populer nomor 1 pada google playstore dengan kategori news serta sudah banyak pengguna yang telah memberikan opininya kepada aplikasi ini. Dalam aplikasi IDN news terdapat berbagai macam fitur seperti berita, live, dan kuis. Analisis sentimen bertujuan untuk mengetahui tanggapan pengguna terhadap aplikasi apakah cenderung ke sentimen positif atau negatif.

3.2 Alur Penelitian



Gambar 3. 1 Alur Penelitian

Pada tahapan Business Understanding dilakukan proses pemilihan terkait dengan judul penelitian serta pemilian objek yang ingin diteliti. Penelitian dilakukan pada sentimen analisis, pemelihan objek dilakukan pada aplikasi IDN dengan komparasi algoritma Random Forest, Decision Tree, KNN. Pada tahapan data understanding dilakukan pemahaman data. Data yang digunakan dalam penelitian ini berisi ulasan pengguna di Indonesia terhadap aplikasi IDN yang diambil dari Google Play Store. Pengumpulan data dilakukan melalui teknik scraping. Fokus data penelitian pada opini dalam bahasa Indonesia. Total data yang diambil mencapai 995 ulasan. Data ulasan yang diambil dari tanggal 1 Maret 2022 sampai 31 Maret 2024. Data yang telah didapatkan akan disimpan dalam format CSV

Setelah mendapatkan data, langkah selanjutnya adalah langkah preprocessing data menggunakan Python. Pada tahap preprocessing data ini dilakukan beberapa proses yaitu:

1. Case folding : Merubah seleluruh huruf kapital pada ulasan
2. Normalization : Mengubah kata singkatan menjadi makna sebenarnya
3. Stopword removal : Menghapus kata yang tidak diperlukan
4. Stemming : Mengubah kata menjadi kata dasar

Data yang telah dibersikan kemudian akan di berikan pembobotan menggunakan *TF-IDF*. Pembobotan dilakukan dengan nilai *N-Gram* 1,1 artinya pemberian bobot akan dilakukan pada setiap kata. Selanjutnya adalah melakukan pelabelan data berdasarkan *score* atau *rating*. Proses belalabelan di bentuk menjadi 2 label yaitu label positif dan label negatif. berdasarkan rating yang telah didapatkan dilakukan pembagian yaitu apabila rating dari ulasan merupakan 4 dan 5 maka akan dilakukan pelabelan positif, jika ulasan memiliki rating 1-3 maka akan masuk ke dalam pelabelan negatif. tidak terdapat label netral karena tidak memiliki pengaruh bobot dalam penelitian. Kemudian data yang telah di berikan bobot akan di seleksi mengguanakn *Chi-Square*. Jumlah fitur yang diseleksi dengan nilai K sebesar 1500 fiur.

Pada proses selanjutnya akan dilakukan modeling dengan menggunakan tiga algoritma yaitu Random Forest, Decision Tree, KNN. Pada proses modeling dilakukan tiga skenario pembagian data test dan data uji dengan pembagian 80:20, 70:30, 60:40. Tahapan modeling dilakukan untuk mendapatkan hasil akurasi, precision, recall, dan f1_score yang didapatkan dengan perhitungan *confusion matrix*. Setalah melakukan proses pemodelan selanjutnya adalah dengan melaukan evaluasi kinerja seluruh algortima. Evaluasi dilakukan dengan membandingkan hasil kinerja algoritma berdasarkan masing masing pembagian skenario data uji. Pada tahapan *deployment* yaitu menyimpan hasil pemodelan dalam bentuk *website* untuk dapat mentukan sentimen positif atau negatif ulasan IDN.

3.3 Metode Penelitian

Pada penelitian ini algoritma yang digunakan merupakan algoritma klasifikasi. Algoritma klasifikasi adalah sebuah metode yang dapat mengidentifikasi berdasarkan kelompok yang ada. Algoritma yang digunakan pada penelitian ini adalah *Random Forest, Decision Tree, KNN*. Alasan menggunakan algoritma tersebut adalah untuk membandingkan performa antar algoritma dan membandingkan dengan hasil akurasi penelitian terdahulu. Pembagian skenario data latih dan data uji dilakukan dengan tiga skenario yaitu: 80:20, 70:30, 60:40. Pembagian data latih dan data uji yang berbeda dilakukan untuk mengetahui tingkat konsistensi algoritma serta kemampuan algortima dalam menangani data yang lebih bessar [9]. Pada algortima KNN menggunakan nilai tetangga terdekat 5 dan 7 pada setiap data uji 20%, 30%, 40. Pembagian nilai tetangga terdekat dilakukan untuk mengetahui pengaruh nilai “K” terhadap hasil akurasi. Nilai “K” yang ideal merupakan bilangan ganjil, Pemilihan nilai k didasarkan pada besaran ukuran data semakin banyak informasi yang terdapat data maka nilai k akan semakin besar [44].Semakin besar nilai K yang digunakan akan mengurangi noise pada data [45]. Jumlah data pada penelitian ini sebesar 995 data sehingga nilai k yang digunakan sebesar 5 dan 7 Terdapat kelebihan dan kekurangan terhadap masing-masing algoritma Berikut merupakan kelebihan dan kekurangan algoritma yang digunakan:

Tabel 3. 1 Kelebihan dan Kekurangan Algoritma

Algoritma	Kelebihan	Kekurangan
Random Forest	- Dapat menangani dataset yang besar dan dengan banyak fitur - Lebih tahan terhadap overfitting dibandingkan decision tree	- Memerlukan sumber daya komputasi yang besar - Waktu pelatihan lebih lama dibandingkan model lain karena kompleksitasnya
Decision Tree	- Tidak memerlukan scaling fitur - Cepat dalam membuat keputusan/prediksi	- Rentan terhadap overfitting, terutama pada pohon yang besar - Tidak efisien untuk dataset yang besar
K-Nearest Neighbor	- Performa yang baik pada data yang terdistribusi secara seragam	- Sensitif terhadap fitur yang tidak relevan atau memiliki skala berbeda - Kurang efisien pada dataset besar

3.4 Teknik Pengumpulan data

Teknik pengumpulan data dilakukan dengan mengambil review aplikasi IDN pada *Google Play Store* adalah dengan teknik scraping. Data yang diambil adalah dengan mengambil ulasan pengguna aplikasi IDN. Ulasan yang diambil adalah ulasan yang berbahasa indonesia. Teknik pengambilan data scraping dilakukan dengan menggunakan bahasa pemrograman *phyton*. Data yang telah didapatkan kemudian akan disimpan dalam file dengan format *CSV*. Jumlah data ulasan pengguna yang terkumpul adalah 995. Data ulasan aplikasi IDN yang diambil dari tanggal 4 Maret 2022 sampai 22 Maret 2024. Setelah data didapatkan dengan format *CSV* selanjutnya adalah melakukan pelabelan terhadap ulasan pengguna. Pemberian label dilakukan berdasarkan rating dari ulasan tersebut.

3.5 Teknik Analisis Data

Teknik Analisis data menggunakan CRISP DM (*Cross Industry Standard Process for Data Mining)*. CRISP DM membuat data yang dihasilkan memiliki hasil yang maksimal karena dalam Teknik ini memiliki tahapan atau alur dalam pengolahan data, dalam setiap tahapan tersebut akan menghasilkan data yang maksimal.

Gambar 3. 2 CRISP-DM

Berikut tahapan dalam CRISP DM

a. *Business Understanding*

   Proses *Business understanding* merupakan tahapan untuk memahami objek dan target yang ingin dicapai. Tahapan ini diperlukan untuk mempersiapkan seluruh jalanya penelitian seperti *tools*, algoritma,dan lain-lain. Topik utama dalam penelitian ini adalah melakukan sentimen analisis aplikasi IDN dengan komparasi algoritma *Random Forest*, *Decision Tree*, dak *KNN*.

b. *Data Understanding*

   Data understanding adalah memahami data yang telah didapatkan, pada proses ini akan dilakukan tahapan dengan pemberian label “positif” dan “negatif” pada ulasan aplikasi IDN.

c. *Data preparation*

Pada tahapan ketiga adanya persiapan data yang akan diolah, tahapan ini memastikan bahwa data siap untuk digunakan serta memastikan bahwa tidak terdapat data *null*. Pada tahapan ini agar data siap di olah terdapat proses pembersihan data dengan *Case folding*, *Normalization*, *Stopword Removal*, dan *Stemming*. Data yang telah bersih maka akan di berikan pembobotan dengan *TF-IDF* dan *N-Gram*, lalu akan di filter dengan *chi-square*.

d. *Modeling*

Pada tahapan ke empat melakukan modeling dari data yang telah dipersiapkan dengan menggunakan algoritma *Random Forest*, *Decision Tree*, *K-Nearest Neighbor*.

e. *Evaluation*

hasil yang didapat akan dikomparasikan dengan penelitian terdahulu yang dijadikan sebagai pembanding.

f. *Deployment*

Pada tahapan terakhir mendeploy prediksi sentimen aplikasi IDN kedalam *website* menggunakan *streamlit*.