BAB II

LANDASAN TEORI

2.1 Data Mining

Menurut (Pasek dkk., 2022) *Data mining* adalah proses ekstraksi informasi yang berharga dan bermanfaat dari kumpulan data yang besar. Ini melibatkan kombinasi berbagai bidang ilmu, termasuk basis data, *information retrieval*, statistika, algoritma, dan *machine learning*

2.2 Analisis Sentimen

Menurut (Sari & Wibowo, 2019) Analisis Sentimen adalah teknik yang digunakan untuk menggali informasi dari teks yang mengungkapkan opini atau pendapat, dengan tujuan untuk menentukan apakah opini tersebut bersifat positif atau negatif,

2.3 Text Mining

Menurut (Reddy dkk., 2022) *Text Mining* adalah proses yang digunakan untuk menggali pola informasi berharga dan pengetahuan dari sejumlah besar data teks, seperti dokumen, file PDF, dan kutipan teks lainnya. Data teks dapat diekstrak dan dianalisis secara efisien untuk mendapatkan wawasan yang berharga dan memperoleh pemahaman yang lebih mudah dari sumber data yang luas.

2.4 Python

Menurut (Muhammad dkk., 2022) *python* adalah bahasa pemrograman yang dikembangkan oleh *Guido Van Rossum* pada tahun 1989. Bahasa ini didesain dengan fokus pada keterbacaan kode dan sintaks yang sederhana

2.5 Text Preprocessing

Menurut (Reddy dkk., 2022) *Text Preprocessing* adalah tahap awal dari web scraping yang melibatkan analisis dan pengolahan informasi dalam data teks semi terstruktur atau tidak terstruktur. Pada tahap ini, data teks diproses dengan mengidentifikasi dan menganalisis pola serta aturan yang ada dalam teks untuk memperoleh data yang lebih terstruktur dan dapat dipahami dengan lebih mudah, berikut tahapan yang dilakukan.

1. Case Folding

Tahap *Case Folding* adalah proses mengubah suatu kata atau kalimat menjadi huruf kecil dengan tujuan untuk mempermudah dalam melakukan pemrosesan data.

2. Cleansing

Tahap *Cleansing* adalah proses membersihkan kata atau kalimat dari tanda baca atau kata-kata yang tidak perlukan dalam pemrosesan data seperti menghilangkan spasi berlebih, menghilangkan tagar, menghilangkan link, menghilangkan emoji dan lain-lain.

3. *Tokenizing*

Pada tahap *Tokenizing* adalah proses memecah kalimat maupun paragraph menjadi bagian-bagian kata yang disebut tokenizing.

4. *Stopword*

Tahap *stopword* adalah proses menghilangkan kata-kata yang dianggap kurang penting berdasarkan kategori yang sudah dibuat, proses ini dibuat dengan tujuan agar data menjadi lebih relevan.

5. *Normalization*

Tahap *normalization* adalah proses merubah kalimat pada data menjadi kalimat baku berdasarkan kamus besar bahasa Indonesia, tahap ini juga melakukan pengubahan kata *slang* (gaul) menjadi kata baku.

6. *Stemming*

Tahap *Stemming* adalah proses mengubah kata-kata menjadi kata dasar dengan menghilangkan imbuhan yang menempel pada kata.

2.6 Algoritma *Naïve Bayes*

Menurut (Yusuf dkk., 2020) Algoritma *Naïve Bayes* adalah algoritma klasifikasi yang digunakan untuk kalkulasi probabilitas dengan melakukan penjumlahan kepada frekuensi dan juga kombinasi nilai dari dataset.

2.7 Algoritma *Support Vector Machine*

Menurut (Petiwi dkk., 2022) *Support Vector Machine* adalah sebuah algoritma klasifikasi yang berguna dalam memproses data teks. *Support Vector Machine* menggunakan metode kernel linear untuk mengelompokkan data ke dalam kelas-kelas yang berbeda. Penerapan kernel membantu dalam menggambarkan data teks dari dimensi yang lebih kecil ke dimensi yang lebih besar, sehingga memungkinkan SVM untuk memahami pola dan hubungan yang kompleks di dalam data teks.

2.8 TF-IDF (Term Frequency) – (Inverse Document Frequency)

Menurut (Nur Akbar dkk., 2022) *TF-IDF* (*Term Frequency-Inverse Document Frequency*) adalah metode yang digunakan untuk memberikan bobot pada kata-kata dalam dokumen berdasarkan frekuensi kemunculannya dalam

dokumen tersebut dan sejauh mana kata tersebut unik atau jarang muncul dalam seluruh koleksi dokumen. Berikut rumus menghitung bobot kata menggunakan *TF-IDF* adalah:

1. TF (*Term Frequency*)

Term Frequency digunakan untuk mengukur seberapa sering kata tertentu muncul *I* dalam dokumen *j* dengan membagi jumlah kemunculan kata tersebut dengan jumlah kata total dalam dokumen *j*. Bobot kata t pada dokumen diberikan dengan:

$$tf = \frac{f_d(i)}{\max f_d(j)\ j}$$

2. IDF (*Inverse Document Frequency*)

IDF adalah perhitungan jumlah kemunculan suatu kata pada setiap dokumen (*df*) dengan total dokumen, yang kemudian akan dilakukan logaritma:

$$idf = \log\left(\frac{n}{df_i}\right)$$

3. **TF-IDF**

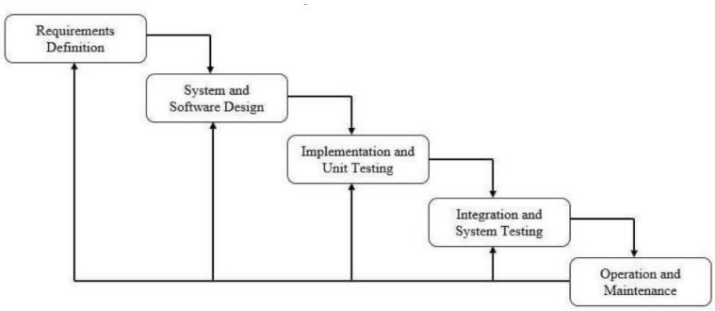
TF-IDF adalah perhitungan nilai *TF* dikalikan dengan *IDF*, berikut ini adalah rumus perhitungan tfidf:

$$w_{ij} = tf.idf$$

2.9 Software Development Life Cycle (SDLC)

Menurut (Safri Irawansyah dkk., t.t.) SDLC adalah salah satu metode pengembangan sistem yang memiliki pendekatan linear dan berurutan. SDLC

Waterfall merupakan kerangka terstruktur berisi tahapan proses – proses sistem dikembangkan, Berikut adalah tahapan dari model waterfall:



Gambar 2. 1 SDLC Diagram Waterfall

1. **Analisis**

   Tahap ini merupakan tahapan yang berisi analisis kebutuhan yang dibutuhkan oleh aplikasi kedepannya. Tahapan ini meliputi pengumpulan data pada sosial media berupa ulasan.

2. **Desain**

   Tahapan ini mengenai rancangan aplikasi yang akan dibuat, seperti pembuatan struktur use case diagram, activity diagram, rancangan tampilan aplikasi.

3. **Implementasi**

   Tahap ini merupakan tentang pembuatan aplikasi meliputi proses implementasi berdasarkan sistem yang sudah diracang pada tahap sebelumnya menjadi sebuah aplikasi melalui tahap coding.

4. **Pengujian**

   Tahap ini meliputi kesesuaian aplikasi agar berjalan sesuai prosedur dan rancangan. Pengujian aplikasi berupa pengujian fungsional.

2.10 Cross-Industry Standard Process for Data Mining (CRISP-DM)

Menurut (Nur Akbar dkk., 2022) *CRISP-DM* adalah salah satu metode yang dapat diterapkan ke dalam strategi pemecahan masalah umum serta metodologi yang menyediakan standar baku untuk data mining. dimana data akan diproses dengan melakukan beberapa tahapan seperti berikut:

1. ***Business Understanding***

   Tahap *Bussiness Understanding* adalah tahap awal dalam proses data mining yang melibatkan penentuan permasalahan yang akan diselesaikan. Dalam kasus ini peneliti ingin mengetahui pandangan pelanggan tentang layanan *IndiHome* dan mengidentifikasi sentimen terkait pengalaman mereka.

2. ***Data Understanding***

   Tahap *Data Understanding* memiliki beberapa bagian yang sangat penting salah satunya adalah mengumpulkan data awal yang relevan pada kasus ini adalah sentimen *IndiHome* pada sosial media lalu mengetahui jumlah data yang akan dilakukan proses data mining.

3. ***Data Preparation***

   Tahap *Data Preparation* melibatkan data mentah yang bertujuan untuk proses data mining, pada tahapan ini juga terjadi proses pemilihan tabel dan kolom yang akan ditransformasikan kedalam data baru.

4. ***Modeling***

   Pada tahap modeling adalah tahapan perhitungan dengan algoritma *Support Vector Machine* dan *Naïve Bayes* untuk melakukan analisis

sentimen dan klasifikasi teks. Selain itu diperlukan untuk memilih tools yang digunakan seperti *Scikit-learn* atau *NLTK*, dan menentukan kernel SVM pada kasus ini menggunakan karnel *linear* dengan bahasa pemrogramman python.

5. Evaluation

Pada tahap *evaluation* dilakukan proses menampilkan hasil akurasi yang diperoleh setelah proses modeling. Tahapan ini juga dapat dilakukan penyesuaian dan pengoptimalan model yang telah dibangun agar dapat menghasilkan pemahaman yang lebih baik lagi tentang data .

6. *Deployment*

Pada tahap ini, model yang telah dikembangkan akan diterapkan dan diuji secara menyeluruh untuk memastikan kesiapan dan kinerjanya dalam menghadapi data baru serta diimplementasikan kedalam lingkungan produksi.

2.11 Confusion Matrix

Confusion Matrix (Aldean dkk., 2022) adalah cara untuk menampilkan hasil akurasi dari model yang telah dibuat. Ini adalah metode yang digunakan untuk meringkas kinerja model dalam melakukan klasifikasi objek berdasarkan berapa banyak kategori yang diklasifikasikan dengan benar.-

Tabel 2. 1 *Confusion Matrix*

Confusion Matrix	*Predicted Positive*	*Predicted Negative*
Actual Positive	TP	TN
Actual Negative	FP	FN

Penjelasan:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

TP = *True Positive* (Jumlah data positif yang terklasifikasi dengan benar)

TN = *True Negative* (Jumlah data negatif yang terklasifikasi dengan benar)

FP = *False Positif* (Jumlah data positif yang terklasifikasi dengan salah)

FN = *False Negative* (Jumlah data negatif yang terklasifikasi dengan salah)

Precision merupakan sebuah definisi rasio antara *True Positive* dan *Total Positive*, *Recall* adalah ukuran yang menggambarkan rasio antara *True Positive* dan *Total Aktual Positive*,

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 = 2.\frac{Precision\ .Recall}{Precision + Recall} = \frac{2.TP}{2.TP + FN + FP}$$

2.12 Crawling Data

Menurut (Nur Akbar dkk., 2022) *Crawling data* adalah proses pengumpulan informasi dari berbagai sumber seperti platform media sosial, atau forum, dengan tujuan mengambil ulasan dan komentar yang memuat sentimen tertentu yang diberikan oleh pengguna terkait suatu produk, layanan, atau topik tertentu.

2.13 Streamlit

Dikutip dari (Streamlit Docs, 2023), *Streamlit* merupakan *library* sekaligus layanan *hosting* yang ada di bahasa pemrograman *Python*, layanan ini mudah untuk digunakan dengan dokumentasi lengkap dan mudah untuk *deploy machine learning* maupun *data science.*

2.14 Deta Base (Deta Space)

Dikutip dari (Deta Space, 2023), "*Deta Base is a fully-managed, fast, scalable and secure NoSQL Database with a focus on end-User simplicity*.".

Deta Base adalah sebuah *Database NoSQL* yang sepenuhnya dikelola, cepat, skalabel, dan aman dengan fokus pada kesederhanaan pengguna akhir.

TEKNOLOGI INFORMASI

UNIVERSITAS DARMA PERSADA