BAB II

TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1 Tinjauan Pustaka

Penelitian ini menggunakan beberapa sumber pustaka yang berhubungan dengan kasus ataupun metode yang akan di teliti, diantaranya yaitu:

Salsabila et al. (2023) melakukan penelitiann dengan judul analisis sentimen penggun *Twitter* terhadap fitur gratis ongkos kirim pada aplikasi *Shopee* Indonesia menggunakan algoritma *K-Nearest Neighbor*. Pada penelitian ini data yang telah diklasifikasikan menghasilkan sentimen positif sebanyak 130 data dan sentimen negatif sebanyak 1.045 data. Dengan pembagian data 80:20, penelitian ini menghasilkan nilai akurasi sebesar 91%. Maka dari itu sentimen pengguna *Twitter* terhadap fitur gratis ongkos kirim ada aplikasi *Shopee* diketahui lebih banyak mengandung sentimen negatif.

Kemudian penelitian yang dilakukan oleh Dwi Fasnuari et al. (2022) dengan judul penerapan algoritma *K-Nearest Neighbor* untuk klasifikasi penyakit diabetes melitus studi kasus: warga desa Jati Tengah. Penelitian ini menggunakan 8 variabel yaitu mudah haus, berat badan turun meskipun konsumsi makanan teratur, tekanan darah tinggi, terdapat riwayat diabetes dalam keluarga, luka yang sulit sembuh, buang sering air dimalam hari, hasil cek gula darah dan usia. Data yang digunakan sebanyak 108 data *training* dan 27 data *testing* menghasilkan akurasi 93% pada K=9, presisi 100% dan *FI-Score* 75%. Dengan tingkat akurasi sebesar 93% maka penelitian ini dinilai telah berhasil menerapkan metode KNN untuk melakukan klasifikasi terhadap penyakit diabetes melitus.

Diwandanu et al. (2023) dengan judul analisis sentimen terhadap *Tweet* maxim pada *Twitter* menggunakan *R programming* dan *K-Nearest Neighbor*. Pada penelitian ini melakukan 3 skema, skema satu menggunakan 80% data latih dan 20% data uji, skema dua menggunakan 75% data latih dan 25% data uji, skema 3 menggunakan 70% data latih dan 30% data uji. Ketiga skema menggunakan k yang berbada yaitu k1 sampai dengan k10. Hasil akurasi terbaik didapatkan saat menggunakan data latih 80% sebanyak 702 data dan data uji 29 & sebanyak 175 data dengan k1 yaitu 95,43%.

Prasetyo et al. (2023) dalam judul Analisis sentimen relokasi Ibukota Nusatra Menggunakan Algoritma *Naïve Bayes* dan KNN. Pada penelitian ini menghasilkan hasil komparasi kinerja metode tersebut bahwa metode *Neive Bayes* dapat memberikan akurasi analisis sentimen sebesar 82.27%, nilai *Precision* sebesar 86.36% dan nilai *Recall* sebesar 76,93%. Kinerja metode KNN juga menyajikan hasil dari analisis dengan tingkat akurasi sebesar 88,12%, *Precision* sebesar 93,98% dan nilai *recall* sebesar 81.53%. berdasarkan analisis tersebut, proses analisis menggunakan metode KNN mengungguli NB dalam mengukur sentimen terhadap pemindahan Ibu Kota Nusantara.

Fikry et al. (2023) dengan judul klasifikasi sentimen masyarakat di *Twitter* terhadap kenaikan BBM dengan metode *K-Nearest Neighbor*. Pada penelitian ini penerapan metode *K-Nearest Neighbor (K-NN)*, *Feature Weighting (TF-IDF)*, dan *Feature Selection (Threshold)* akan dilakukan implementasi menggunakan tools *google collab*. Hasil pengujian yang didapatkan dari pengujian metode KNN dengan *confusion matrix* pada 10 nilai K yang berbeda dengan mekanisme

perbandingan yang digunakan 70:30, 80:20, dan 90:10 diperoleh akurasi paling besar 83,3% pada k=13 dan k=15 untuk perbandingan data *training* dan *testing* 90:10.

Tabel 2. 1 Penelitian Terdahulu

Penulis	Topik	Metode	Keterangan
Salsabila et al. (2023)	Analisis sentimen pengguna *twitter* terhadap fitur gratis ongkos kirim pada aplikasi *Shopee* Indonesia	*K-Nearest Neighbor*	Menghasilkan nilai akurasi sebesar 91%. Maka dari itu sentiment pengguna *Twitter* terhadap fitur gratis ongkos kirim pada aplikasi *Shopee* diketahui lebih banyak mengandung sentimen negatif
Fasnuari et al. (2022)	Klasifikasi penyakit diabetes melitus	*K-Nearest Neighbor*	Klasifikasi dengan metode Data yang digunakan sebanyak 108 data trainening dan 27 data testing menghasilkan akurasi 93% pada K=9, presisi 100% dan *FI-Score* 75%. Dengan tingkat akurasi sebesar 93%
Diwandanu & wisudawati (2023)	Analisis sentiment terhadap twit *Maxim* pada *Twitter*	*K-Nearest Neighbor*	Hasil akurasi terbaik didapatkan saat menggunakan data latih 80% sebanyak 702 data dan data uji 29& sebanyak 175 data dengan k1 yaitu 95,43%
Prasetyo et al. (2023)	Analisis sentimen relokasi Ibukota Nusantara	*Naïve Bayes* dan *K-Nearest Neighbor*	Menghasilkan hasil komparasi kinerja metode tersebut bahwa metode *Neive Bayes* dapat memberikan akurasi analisis sentimen sebesar 82.27%, nilai *Precision* sebesar 86.36% dan nilai *Recall* sebesar 76,93%. Kinerja metode KNN juga menyajikan hasil dari analisis dengan tingkat akurasi sebesar 88,12%, *Precision* sebesar 9 3,98% dan nilai *recall* sebesar 81.53%.

Tabel 2. 1 (Lanjutan)

Penulis	Topik	Metode	Keterangan
Arista et al. (2023)	Klasifikasi Sentimen Masyarakat di *Twitter* terhadap kenaikan harga BBM	*K-Nearest Neighbor*	Hasil pengujian yang didapatkan dari pengujian metode KNN dengan *confusion matrix* pada 10 nilai K yang berbeda dengan mekanisme perbandingan yang digunakan 70:30, 80:20, dan 90:10 diperoleh akurasi paling besar 83,3% pada k=13 dan k=15 untuk perbandingan data *training* dan *testing* 90:10.
Zuliani (2024)	Analisis sentimen ulasan pengguna aplikasi *X (Twitter)* di *Playstore*	*K-Nearest Neighbor*	Klasifikasi dengan metode KNN pada penelitian ini menggunakan 2 perbandingan pembagian data 90:10 dan 80:10. Nilai K yang digunakan merupakan nilai K terbaik berdasarkan Range.

2.2 Landasan Teori

Landasan teori merupakan sekumpulan bahan pustaka yang berhubungan dengan permasalahan yang akan diteliti. Dengan adanya landasan teori diharapkan dapat dapat digunakan untuk memecahkan permasalahan yang diangkat dalam penelitian ini yaitu: Penerapan metode K-NN untuk analisis sentimen ulasan pengguna aplikasi *X* di *Playstore.*

2.2.1 Data mining

Data mining adalah studi tentang mengumpulkan *(collection),* pembersihan *(cleaning),* pengolahan *(processing)* menganalisis *(analyzing),* dan memperoleh *(gaining)* wawasan atau pandangan yang berguna dari data tersebut. Ada variasi yang luas dalam domain masalah, aplikasi, formulasi dan representasi data yang dihadapi dalam aplikasi nyata. Data mining adalah istilah luas paying yang

digunakan untuk menggambarkan aspek-aspek berbeda dari pengolahan data (data *processing*). Proses data mining memiliki beberapa fase seperti pembersihan data (data *cleaning*), ekstrak fitur (*feature extraction*), dan desain algoritma (*algorithmic design*). (Indra et al. 2017)

2.2.2 K-Nearest Neighbor

K-Nearest Neighbor atau biasa disebut dengan KNN adalah salah satu metode paling sederhana untuk memecahkan masalah klasifikasi. Algoritma sering digunakan untuk klasifikasi teks dan data. Pada metode ini dilakukan klasifikasi terhadap obyek berdasarkan data yang jaraknya paling dekat dengan obyek tersebut (Nurjanah et al. 2017).

Penggunaan metode KNN dalam klasifikasi teks dapat menghasilkan nilai yang lebih optimal dengan menerapkan rumus *cosine similarity* untuk memberikan bobot pada setiap kata dalam dokumen teks yang sedang diproses. Metode *cosine similarity* merupakan metode untuk menghitung kesamaan antara dua objek yang dinyatakan dalam dua *vector* objek. Ketika lebih banyak karaktek muncul dalam sebuah dokumen, tingkat kesamaan dapat ditentukan meningkat. Persamaan dari *cosine similarity* ditunjukan pada persamaan 2.1.

$$CosSim(q, dj) = \frac{dj.q}{|d_j|.|q|} = \frac{\sum_{i}^{t}=1\ (w_{iq})}{\sqrt{\sum_{i}^{t}=1\ w_{ij}2.}\sqrt{\sum_{i}^{t}=1\ w_{ij}2.}} \qquad (2.1)$$

Keterangan:

CosSim (q,d_j) Nilai kemiripan antara dokumen uji (q) dengan dokumen latih ke j (d).

t = jumlah *term* (kata)

d = dokumen

q = kata kunci *(query*)

W_{ij} = bobot *term* (kata) ke 1 pada dokumen latih j

W_{iq} = bobot *term* (kata) ke 1 pada dokumen uji q

2.2.3 Confusion Matrix

Confusion Matrix merupakan tahap evaluasi menggunakan perhitungan akurasi, presisi dan *recall* dari hasil klasifikasi. (Khoirudin et al. 2018). Tabel *confusion matrix* dapat dilihat pada Tabel 2.2.

Tabel 2. 2 Tabel *Confusion Matrix Multi Class*

Realita	Nilai Prediksi			
	Kelas 1	Kelas 2	Kelas 3	Total
Kelas 1	*True Positive*	*Error*	*Error*	Total Kelas 1
Kelas 2	*Error*	*True Negative*	*Error*	Total kelas 2
Kelas 3	*Error*	*Error*	*True Positive*	Total kelas 3

Keterangan:

True positive, merupakan data positif yang diprediksi benar.

True Negative, merupakan datat negatif yang diprediksi benar.

False Positive (Error), merupakan data negatif namun diprediksi sebagai data positif.

False Negative (Error), merupakan data positif namun diprediksi sebagai data negatif.

Berdasarkan tabel 2.2. untuk menghitung akurasi menggunakan persamaan 2.2, presisi dengan persamaan 2.3, *recall* dengan persamaan 2.4 dan *F1-Score* dengan persamaan 2.5. Perhitungan yaitu sebagai berikut:

1. Akurasi

Merupakan ukuran hasil kerja yang memberikan tingkat keakuratan dari seluruh model.

$$Accuracy = \frac{TP(kelas\ 1)+TP(Kelas\ 2)+TP(Kelas\ 3)}{Total(kelas\ 1)+Total(Kelas\ 2)+Total(Kelas\ 3)} \qquad (2.2)$$

2. Precission

Precssion merupakan tingkat ketepatan benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif.

$$Precision = \frac{TP(Kelas\ i)}{Prediksi\ (Kelas\ i)} \qquad (2.3)$$

3. Recall

Recall merupakan prediksi benar positif dibandingkan dengan keseluruhan data yang sebenarnya positif.

$$Recall\ = \frac{TP(Kelas\ i)}{Total\ (Kelas\ i)} \qquad (2.4)$$

4. F1-Score

F1-Score merupakan rata-rata harmonik dari *precission* dan *recall* yang menghitung kedua *matrix* tersebut. *F1-Score* Menghitung keseimbangan antara *precission* dan *recall.*

$$F1 - Score = \frac{2 \times precision \times recall}{precision+recall} \qquad (2.5)$$

2.2.4 Analisis sentimen

Analisis sentimen merupakan analisis terkait opini masyarakat yang diungkapkan dalam bentuk teks. Opini publik memiliki beberapa bentuk penilaian, evaluasi, emosi, kepercayaam, kepuasan. Analisis sentimen menghasilkan sentimen positif, sentimen negatif dan sentimen netral (Larasati et al. 2022).. Kemudian analisis sentimen memiliki 4 jenis:

1. Analisis sentimen bertingkat, digunakan untuk mengartikan bintang dalam ulasan. Bintang 5 diartikan positif dan bintang 1 diartikan negatif.
2. Mengetahui emosi, digunakan untuk mengetahui emosi misalnya rasa Bahagia, marah, sedih dan frustasi.
3. Berlandas fitur digunakan untuk mendeteksi fitur yang mempunyai sentimen positif, negatif dan netral.
4. Multibahasa, digunakan untuk mengetahui bahasa dalam sebuah teks menggunakan bahasa yang dipilih.

2.2.5 TF-IDF

TF-IDF *(Term Frequency- Inverrs Document Frequency)* ialah metode untuk memberikan bobot pada kata yang terdapat pada dokumen. TF-IDF terdapat dua tahap yaitu TF dan IDF. Tf merupakan jumlah kemunculan kata pada dokumen, semakin banyak kata tersebut muncul maka semakin besar pula nilai TF. Sedangkan IDF adalah jumlah seluruh dokumen yang mengandung kata tersebut jika kata terebut jarang muncul pada suatu dokumen maka IDF lebih besar dari kata yang sering muncul (Assidyk et al. 2020).

Nilai IDF didapatkan dengan persamaan 2.6 berikut:

$$tf_{td}idf_t = tf_{td} * \log\left(\frac{N}{dft}\right) \qquad (2.6)$$

Dimana:

$tfx\ idf$: Bobot total dari kata t

tf_{td}: Jumlah kemunculan kata t dalam suatu dokumen

N: Total dokumen

df_t: Jumlah dari seluruh dokumen yang mengandung kata t

2.2.6 *Python*

Python adalah bahasa pemrograman interpretatif multiguna dengan filosofi perancangan yang berfokus pada tingkat keterbacaan kode. Tujuan khusus Bahasa pemrograman *Python* ini dikembangkan untuk membuat *source code* yang mudah untuk dibaca. *Python* memiliki *library* yang cukup lengkap sehingga memungkinkan programmer untuk membuat aplikasi yang terbaru dengan menggunakan *source code* yang tampak sederhana. *Python* termasuk kedalam 3 bahasa pemrograman yang telah digunakan selama beberapa tahun belakangan ini. *Python* juga memiliki beberapa *library* yang familiar dalam data science ataupun *mechine learning*, *library* tersebut yaitu: *Scikit-Learn*, *TensorFlow* dan *PyTorch* (Al Farobi, 2021).

2.2.7 Sastrawi

Sastrawi adalah *library python* untuk mengubah kata-kata kedalam Bahasa Indonesia dari bentuk infleksi menjadi kata dasar. Sastrawi digunakan pada tahap *preprocessing stopword removal* dan *stemming* (Cahya, 2023).

2.2.8 Scikit learn

Scikit-learn adalah sebuah *library mechine learning* bersifat *open source*. *Scikit-learn* diterapkan untuk *supervised learning* atau *unsupervised learning*. Pada *library* ini terdapat fasilitas untuk menyesuaikan model, melakukan *pre-processing* data, memilih model dan menyediakan *matrix* untuk mengevaluasi kinerja model (Cahya, 2023).

2.2.9 Web Scrapping

Dalam Parasati et al. (2020), Vargiu dan Urru menyatakan bahwa *web scrapping* merupakan cara untuk mendapatkan data atau informasi dari situs web yang dilakukan secara otomatis. *Web scrapping* bertujuan untuk mencari informasi dari situs web yang berbeda dan tidak terstuktur lalu diubah menjadi bentuk yang terstruktur berbentuk *spreadsheet*, basis data atau *comma separated values (CSV).*