BAB II

LANDASAN TEORI

2.1 Text Mining

Text Mining adalah proses untuk memperoleh informasi berkualitas tinggi dari teks. Informasi berkualitas tinggi biasanya diperoleh karena memperhatikan pola dan tren melalui cara mempelajari pola statistik. Text mining biasanya melibatkan proses penataan teks input (biasanya parsing, bersama dengan penambahan beberapa fitur linguistik dan penghapusan kata, dan selanjutnya menyisipkannya kedalam database), menurunkan pola dalam data yang terstrTextur, dan akhirnya mengevaluasi dan mengintrepetasi output. 'Kualitas tinggi' di text mining biasanya mengacu pada beberapa kombinasi dari relevansi, kebaruan, dan kemenarikan. Text mining biasanya termasuk kategorisasi teks, teks klustering, ekstraksi konsep/entitas, produk taksonomi granular, analisis sentimen, merangkum dokumen, dan pemodelan hubungan entitas (misalnya, mempelajari hubungan antar entitas). [1][2]

2.2 TAPERA (Tabungan Perumahan Rakyat)

Tabungan Perumahan Rakyat (Tapera) adalah penyimpanan yang dilakukan oleh peserta secara periodik dalam jangka waktu tertentu yang hanya dapat dimanfaatkan untuk pembiayaan perumahan dan/atau dikembalikan berikut hasil pemupukannya setelah kepesertaan berakhir. Menghimpun dan menyediakan dana murah jangka panjang yang berkelanjutan untuk pembiayaan perumahan dalam rangka memenuhi kebutuhan rumah yang layak dan terjangkau bagi Peserta.[3]

2.3 Media Sosial

Media sosial (sering disalahtuliskan sebagai sosial media) adalah sebuah media daring, dengan para penggunanya bisa dengan mudah berpartisipasi, berbagi, dan menciptakan isi blog, jejaring sosial, wiki, forum dan dunia virtual. Blog, jejaring

sosial, dan wiki merupakan bentuk media sosial yang paling umum digunakan oleh masyarakat di seluruh dunia. [4]

2.4 Sentimen Analisis

Analisis sentimen, juga disebut penambangan opini, adalah bidang studi itu menganalisis pendapat, sentimen, evaluasi, penilaian, sikap, dan emosi terhadap entitas seperti produk, layanan, organisasi, individu, masalah, peristiwa, topik, dan atributnya. Itu mewakili besar ruang masalah. Ada juga banyak nama dan tugas yang sedikit berbeda, misalnya, analisis sentimen, penambangan opini, ekstraksi pendapat, sentimen penambangan, analisis subjektivitas, analisis pengaruh, analisis emosi, peninjauan penambangan, dan lain-lainnya. Namun, sekarang semuanya berada di bawah payung analisis sentimen atau penambangan opini. Sementara di industri, analisis sentimen jangka lebih umum digunakan, tetapi di akademisi baik analisis sentimen dan penambangan opini sering digunakan.[5]

2.5 Twitter Rest API

Twitter adalah jejaring sosial populer di mana pengguna dapat berbagi pesan singkat seperti SMS (*Short Message Service*) yang disebut *tweet*. Pengguna berbagi pemikiran, tautan, dan gambar di *Twitter*, jurnalis mengomentari secara langsung, perusahaan mempromosikan produk, dan terlibat dengan pelanggan. Berbagai cara dalam menggunakan *Twitter*, dengan 500 jutaan *tweet* per hari, ada banyak data untuk dianalisis dan dimainkan. *Twitter Rest API* menyediakan akses terprogram untuk membaca dan menulis data *Twitter*. Buat *tweet* baru, baca profil penulis dan data pengikut, dan banyak lagi. *REST API* mengidentifikasi aplikasi *Twitter* dan pengguna yang menggunakan *OAuth*.[6]

2.6 Twitter

Twitter adalah layanan bagi teman, keluarga, dan teman sekerja untuk berkomunikasi dan tetap terhubung melalui pertukaran pesan yang cepat dan sering. Pengguna memposting *Tweet*, yang dapat berisi foto, video, tautan, dan teks. Pesan

ini diposting ke profil Anda, terkirim ke pengikut, dan dapat dicari di pencarian Twitter.[7]

2.7 Tweet

Tweet adalah setiap pesan yang diposting ke Twitter dan dapat berisi foto, video, tautan, serta teks. Klik atau sentuh tombol Tweet untuk memposting pembaruan ke profil Anda.

2.8 Orange Data Mining

Orange adalah software open source untuk pengolahan *Data Analytics / Data Mining*. Dibandingkan dengan software *Data Mining* lainnya, *Orange* unggul dalam hal visualisasi atau yang biasa kita sebut *visual programming*. *Orange* menyediakan banyak *widget* yang kita letakkan pada *canvas / drawing board* kemudian kita hubungkan dengan *widget widget* lainnya. Dengan media *canvas* ini maka akan memudahkan pemakai bermain dengan data dan melakukan proses *data analytics* secara intuitif. Selain *widget*, *orange* juga menyediakan beberapa add-on / modul untuk problem pada domain domain tertentu seperti *Text Mining / Text Analytics, Bioinformatics, Network Data / Social Network, Model Maps, Prototypes Process*, dan lain lain. [8][9]

2.9 K-Nearest Neighbor

Metode tunggal yang membaca titik data dikategorikan sebagai pencari data atau objek yang lebih sering muncul pada proses klasifikasi. Metode ini bisa digunakan juga pada klasifikasi umum yang sering digunakan yaitu multiclass. Tetapi metode ini dalam menentukan dan menemukan titik terdekat dengan query dari database masih sangat rendah[10]

k-Nearest Neighbors (k-NN) adalah algoritma machine learning yang digunakan untuk klasifikasi dan regresi. Algoritma ini bekerja dengan mencari k tetangga terdekat dari data baru berdasarkan jarak, biasanya menggunakan jarak Euclidean. Dalam klasifikasi, data baru diberi label kelas yang paling sering muncul di antara

tetangga terdekatnya, sedangkan dalam regresi, nilai prediksi diperoleh dari rata-rata nilai tetangga tersebut.

Pemilihan nilai k sangat penting karena mempengaruhi akurasi; k yang terlalu kecil bisa membuat model terlalu sensitif terhadap outlier, sedangkan k yang terlalu besar bisa membuat model kurang responsif terhadap pola lokal. Metode ini sederhana dan fleksibel, namun bisa kurang efisien pada dataset besar dan sensitif terhadap skala fitur.

Algoritma ini bekerja dengan mencari k tetangga terdekat dari data baru berdasarkan jarak, biasanya menggunakan **jarak Euclidean**. Rumus jarak Euclidean antara dua titik $\mathbf{x} = (x_1, x_2, \dots, x_n)$ dan $\mathbf{y} = (y_1, y_2, \dots, y_n)$ dalam ruang n -dimensi adalah:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2.1)$$

Langkah-langkah k -NN:

1. Hitung jarak antara data baru dengan semua data dalam dataset menggunakan rumus di atas.
2. Pilih k tetangga terdekat berdasarkan jarak yang paling kecil.
3. Klasifikasi: Tentukan kelas data baru berdasarkan voting mayoritas dari kelas k tetangga terdekat.
4. Regresi: Prediksi nilai data baru sebagai rata-rata nilai dari k tetangga terdekat.

Pemilihan nilai k penting karena mempengaruhi akurasi; k yang terlalu kecil bisa membuat model terlalu sensitif terhadap outlier, sementara k yang terlalu besar bisa membuat model kurang responsif terhadap pola lokal. Metode ini sederhana dan

fleksibel, namun bisa kurang efisien pada dataset besar dan sensitif terhadap skala fitur.

2.10 Support Vector Machine

Support Vector Machine (SVM) merupakan salah satu algoritma HMS yang digunakan untuk memprediksi regresi dan klasifikasi. Prinsip dasar yang digunakan dalam algoritma ini adalah pengklasifikasi linier yang merupakan kasus klasifikasi yang dapat dipisahkan secara linear dan juga dapat digunakan pada kasus nonlinier dengan memasukkan konsep kernel ruang berdimensi tinggi dengan mencari hyperplanes yang dapat memaksimalkan jarak antar kelas. [11]

SVM bertujuan untuk menemukan hyperplane yang memaksimalkan margin, yaitu jarak terdekat antara titik data dari setiap kelas ke hyperplane tersebut. Titik data yang terletak pada margin disebut support vectors, dan mereka adalah elemen penting yang menentukan posisi hyperplane

Misalkan kita memiliki dataset yang terdiri dari n titik data, di mana setiap titik data memiliki label y_i yang bernilai $+1$ atau -1 , dan x_i adalah vektor fitur.

Hyperplane dapat didefinisikan sebagai:

$$w \cdot x - b = 0$$

Di mana:

- w adalah vektor bobot.
- x adalah vektor fitur.
- b adalah bias atau intercept.

Margin didefinisikan sebagai jarak antara hyperplane dan titik terdekat dari kedua kelas, dan SVM bertujuan untuk memaksimalkan margin tersebut. Jarak dari titik x_i ke hyperplane adalah:

$$\text{Jarak} = y_i (w \cdot x_i - b) / |w|$$

SVM memaksimalkan margin dengan menyelesaikan masalah optimasi berikut:

$$\min_{\{w, b\}} \frac{1}{2} \|w\|^2$$

Dengan syarat:

$$y_i (w \cdot x_i - b) \geq 1$$

2.11 website

Menurut Yeni Susilowati (2019) Website adalah sejumlah halaman web yang memiliki topik saling terkait antar satu halaman dan halaman lainnya, biasanya ditempatkan pada sebuah server web yang dapat diakses melalui jaringan internet maupun jaringan wilayah lokal (LAN).

2.12 Penelitian terdahulu

Dalam penelitian ini akan digunakan tiga penelitian terdahulu yang nantinya dapat mendukung penelitian, berikut ini merupakan penelitian terdahulu yang diambil yaitu:

Table 1 penelitian terdahulu

Nama Penelitian	Judul	Masalah	Hasil penelitian
Nada Adela 2024	Analisis Sentimen Ulasan Pengguna Aplikasi Seabank Menggunakan Support Vector Machine Dan Naïve Bayes	Masalah yang teridentifikasi dalam skripsi mengenai review pengguna terkait layanan seabank dalam transfer, tabungan dan transaksi investasi	Menunjukkan bahwa hasil penelitian dalam analisis sentimen dengan label netral lebih banyak dibandingkan lainnya dengan menggunakan kedua metode bahwa algoritma svm memiliki akurasi tertinggi dibandingkan naïve bayes

Fitriana 2021	Analisis sentimen opini terhadap vaksin covid-19 pada media sosial twitter menggunakan support vector machine dan naïve bayes	Masalah yang teridentifikasi mengenai opini masyarakat yang tertera di dalam aplikasi twitter terkait vaksin covid-19	Hasil kinerja dari algoritma bayes menunjukkan akurasi sebesar 81,20% recall 79,60% dan presisi 79,40% disisi lain algoritma svm mencapai akurasi sekitar 85% recall 84% dan presisi 83,60%.
Petiwi 2022	Analisis sentimen gofood berdasarkan twitter menggunakan metode naïve bayes dan support vector machine	Masalah yang teridentifikasi mengenai sentimen masyarakat di media sosial twitter terkait pelayanan gofood.	Hasil analisis sentimen masyarakat terkait gofood di twitter menunjukkan bahwa sekitar 92,8% dari sentimen bersifat netral, sementara 5,2% bersifat positif dan 2,0% bersifat negatif. Dalam perbandingan akurasi metode support vector machine mengungguli metode naïve bayes dengan akurasi svm mencapai 98,5% dan naïve bayes 91,5%.
Mahendra	Analisis Sentimen Opini Masyarakat	Masalah yang teridentifikasi	Berdasarkan pengujian yang

2022	Terhadap Fenomena TikTokShop di Indonesia Menggunakan Metode K-Nearest Neighbor berbasis N-gram dengan Seleksi Fitur Information Gain	adalah sentimen masyarakat terkait tiktokshop.	telah dilakukan, didapatkan hasil bahwa nilai terbaik terdapat pada fitur Unigram dan nilai threshold yang digunakan adalah 100%, menghasilkan akurasi sebesar 89%, dengan recall 89%, Precision sebesar 89.00%, dan FMeasure sebesar 89.00%. Berdasarkan temuan tersebut, dapat disimpulkan bahwa dalam menganalisis sentimen opini masyarakat Indonesia terhadap TikTokShop, metode K-Nearest Neighbor (KNN) dengan fitur Unigram dan tanpa seleksi fitur Information Gain memberikan hasil terbaik.
2024	Analisa Sentimen Twitter Terhadap Capres Indonesia	Masalah yang teridentifikasi adalah opini	Penelitian ini memanfaatkan data tweet yang

	2024 Menggunakan Metode KNN	masyarakat yang ada di media sosial twitter mengenai capres.	dikumpulkan dari platform Twitter. Program Python di Google Colabs digunakan untuk pengumpulan data, dan perpustakaan TextBlob digunakan untuk pelabelan data. Calon presiden merupakan kata kunci yang digunakan dalam proses pendataan. Terdapat 1.458 tweet dengan sentimen positif sebesar 40,31 % dan sentimen negatif sebesar 43,46 %. Hasil penataan dengan strategi K-Nearest Neighbor memperoleh ketepatan laju sebesar 90,95%, disusul akurasi sebesar 62,17%, tinjauan sebesar 62,23%, dan Fproporsi sebesar 61,87%. Dalam penelitian ini telah dilakukan analisa terhadap tweet pada twitter untuk mencari
--	-----------------------------------	---	---

			<p>optimasi sentiment analisa positif dan negatif dengan menggunakan program python di google colab dan aplikasi orange data mining. Dari hasil penelitian yang telah dilakukan terbukti bahwa optimasi sentimen analisis dari tweet pada twitter mendapatkan hasil yang baik menggunakan metode K-Nearest Neighbor dengan nilai F1-Score, akurasi, presisi, dan recall lebih tinggi dibandingkan metode lainnya.</p>
--	--	--	---