BAB II
TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

Tinjauan pustaka digunakan sebagai acuan penulis sebagai perbandingan penelitian dan sebagai sumber untuk menambah bahan kajian selama penelitian. Selain itu, landasan teori ditingkatkan oleh penulis menggunakan tinjauan pustaka.

Tabel 2.1 Penelitian Sebelumnya

No	Judul	Peneliti	Metode	Masalah Penelitian	Hasil
1	Klasifikasi sentimen masyarakat terhadap kenaikan harga tiket pesawat pada twitter menggunakan naive bayes (2019)	Agustina Merdeka Raya , Fitri Nurbaiti,Detin Sofia	Algoritma Naive Bayes	sentimen masyarakat terhadap kenaikan harga tiket pesawat	Tingkat akurasi klasifikasi menggunakan Naive Bayes sebesar 90.70%
2	Analisis Sentimen Twitter Terhadap Tokoh Publik Dengan Algoritma Naive Bayes Dan Support Vector Machine(2019)	Tanthy Tawaqalia Widowati, Mujiono Sadikin	Naive Bayes Dan Support Vector Machine	Sentimen Twitter Terhadap Tokoh Publik	Menurut temuan, akurasi Naive Bayes adalah 91,48%, sedangkan SVM adalah 85,47 persen. Skor presisi SVM adalah 90,95%, sedangkan skor presisi Naive Bayes adalah 89,28%. Hasil recall

No	Judul	Peneliti	Metode	Masalah Penelitian	Hasil
					Naive Bayes adalah 91,58 %, sedangkan SVM adalah 76,18 %.
3	Analisis Sentimen Pada Twitter Terhadap UIN Raden Fatah Menggunakan Support Vector Machine (2020)	Gusmelia Testiana, Dian Erlina	Support Vector Machine	Sentimen Pada Twitter Terhadap UIN Raden Fatah	Berdasarkan temuan penelitian, sampel 100 tweet menunjukkan polaritas sentimen Twitter terhadap UIN Raden Fatah Palembang: 89 tweet (89 %) memiliki sentimen positif dan 11 tweet (11%) memiliki sentimen negatif. Akurasi klasifikasi sentimen SVM adalah 70%, dengan presisi rata-rata 20,6%, rata-rata recall 70%, dan rata-rata f-measure 62,7%.
4	Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid-19 Menggunakan Algoritma	Winda Yulita, Eko Dwi Nugroho, Muhammad Habib Algifari	Naïve Bayes Classifier	Opini Masyarakat Tentang Vaksin Covid-19	Hasil analisis menunjukkan bahwa mayoritas tweet (60,3%) positif, sedangkan proporsi tweet netral (34,4%) lebih tinggi dibandingkan proporsi tweet yang menentang

No	Judul	Peneliti	Metode	Masalah Penelitian	Hasil
	Naïve Bayes Classifier (2021)				(5,4%). Akurasi yang dihasilkan adalah 0,93.
5	Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19 (2021)	Brian Laurensz,Eko Sediyono	Naive Bayes Dan Support Vector Machine	Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19	Hasil klasifikasi dari metode Naïve Bayes mendapatkan rata-rata akurasi 85,59%, sedangkan SMV sebesar 84,41%.
6	Analisis Sentimen Vaksin Sinovac Pada Twitter Menggunakan Algoritma Naive Bayes (2021)	Sri Lestari, Sudin Saepudin	Algoritma Naive Bayes	Vaksin Sinovac	Studi ini menemukan bahwa tweet dengan sentimen positif sebanyak 86% dari total, sedangkan tweet dengan sentimen negatif sebanyak 14%. Perhitungan yang dilakukan dengan algoritma Naive Bayes memiliki nilai akurasi sebesar 92,96 persen.

Perbedaan antara penelitian ini dengan penelitian sebelumnya terletak pada objek penelitian. Pada penelitian ini yang menjadi objeknya yaitu sirkuit

Mandalika. Selain itu, dalam penelitian ini dilakukan proses *labeling* secara otomatis menggunakan Textblob serta penelitian ini melakukan perbandingan antara tiga sentimen dan dua sentimen.

2.2 Dasar Teori

Berikut adalah penjelasan mengenai teori yang menjadi landasan penelitian:

2.1.1 Sirkuit Mandalika

Sirkuit mandalika di bangun mulai September 2019 dan di resmikan pada 12 November 2021. Sirkuit dengan nama resmi Pertamina Mandalika Internasional Street Circuit memiliki panjang 4,3 kilometer dengan 19 tikungan [3]. Sirkuit mandalika adalah sirkuit internasional yang telah lama menjadi wacana di Indonesia. Sirkuit yang berada di Lombok Tengah NTB ini menjadi pusat perhatian internasional dan Indonesia khususnya. karena sirkuit yang dibangun dengan menelan biaya 3 milyar dolar Amerika tersebut akan di jadikan sebagai tempat penyelenggaraan ajang balap bergengsi World Superbike di November 2021 dan MotoGP pada tahun 2022 [4].

2.1.2 Twitter

Salah satu platform media sosial di Indonesia dengan basis pengguna tahunan yang cukup besar adalah Twitter. Tercatat pada 2019, pengguna Twitter di Indonesia mencapai 6,3 juta pengguna, dan jumlah tersebut menyumbang 52.000 pengguna jejaring sosial. Twitter banyak digunakan oleh publik, sektor swasta, bahkan pemerintah. Twitter digunakan untuk mengkomunikasikan kebijakan, menyampaikan informasi, menerima informasi, atau memberikan umpan balik tentang suatu masalah [2].

2.1.3 Python

Python adalah bahasa pemrograman berorientasi objek dinamis yang digunakan untuk berbagai proyek pengembangan perangkat lunak [7]. Guido Van Rossum menciptakan python, yang dirilis di Belanda pada tahun 1990. Acara televisi yang disukai Guido, Monty Python's Flying Circus, adalah sumber dari

nama python. Karena sintaks sederhana dan perpustakaan yang luas, python adalah bahasa pemrograman yang populer di industri dan pendidikan [8].

2.1.4 Google colab

Google colab atau Google collaboratory adalah sebuah *tools* yang berbentuk *cloud* dan bersifat gratis yang di sediakan Google. Google Colab dibuat menggunakan lingkungan Jupyter dan mendukung hampir semua *library* [9]. Google Colab menggunakan python sebagai bahasa pemrogramannya dan juga menyediakan dua jenis kemampuan perangkat lunak untuk menjalankan kode program tertulis: GPU dan TPU [10].

2.1.5 TextBlob

Textblob adalah salah satu *library* di pyhton versi 2 dan 3 yang dapat mengelola data teks dan mudah diakses untuk pembuatan prototipe cepat [11]. Alat pemrosesan bahasa alami (NLP) seperti ekstraksi frase kata benda, analisis sentimen, klasifikasi, dan terjemahan tersedia dari *Textblob*. Pembelajaran bahasa alami dilakukan dengan objek Textblob yang dihasilkan, tetapi perpustakaan ini hanya mengenal bahasa Inggris [12].

2.1.6 Crawling

Crawling adalah proses untuk mengumpulkan data dari suatu database dengan cara mengunduh atau mengambil data dari sumber tersebut [6]. Menggunakan Twitter API, *crawling* dapat mengunduh atau mengambil data dari server Twitter. Data yang dikumpulkan adalah data pengguna, dan *tweet* itu sendiri adalah data [11].

2.1.7 Analisis Sentimen

Tujuan dari analisis sentimen, yang merupakan komponen dari penambangan data, adalah untuk mengekstraksi, memproses, dan menganalisis data tekstual dari media sosial tentang subjek, orang, organisasi, peristiwa, atau produk tertentu [13]. Dengan kata lain, analisis sentimen adalah proses pemisahan kalimat teks menjadi dua kelompok: positif dan negatif [14].

2.1.8 Klasifikasi

Proses pengklasifikasian objek data dari sekian banyak kelas yang ada dikenal dengan istilah klasifikasi [15]. Klasifikasi berfungsi dengan membuat model data pelatihan yang ada dan menggunakan model tersebut untuk pengklasifikasi data baru [16].

2.1.9 Pre-processing

Preprocessing adalah tahap untuk merapihkan data teks agar dapat digunakan pada tahap berikutnya. Dengan melakukan proses ini diharapkan menghasilkan data teks menjadi data yang rapih dan memiliki kualitas yang baik. Terdapat beberapa proses *preprocessing* pada penelitian ini yaitu:

1. *Cleaning* adalah proses pembersihan data dari tanda baca, angka, simbol, url dan username.
2. Proses mengubah karakter huruf besar semua data teks menjadi huruf kecil dikenal sebagai *case folding*.
3. *Remove Duplicate,* tujuan dari prosedur ini adalah untuk menghilangkan data yang identik atau duplikat dari setiap teks dalam data.
4. *Tokenizing,* proses mengubah kalimat menjadi token atau kata demi kata.
5. *Stopword Removel* adalah proses pemilihan kata-kata penting setelah dilakukan proses tokenisasi.
6. *Stemming* adalah interaksi yang digunakan untuk mengubah ekspresi suatu bahasa menjadi bentuk dasarnya, yang menghilangkan awalan, imbuhan, atau awalan yang tidak berguna [17].

2.1.10 Naive Bayes

Naive Bayes adalah metode klasifikasi yang didasarkan pada Teorema Bayes dan menggunakan metode statistik dan probabilitas.. Keuntungan dalam menggunakan Naive Bayes adalah hanya menggunakan data *training* yang sedikit untuk menetapkan estimasi parameter yang diperlukan untuk pengklasifikasian [18]. Naive Bayes mengasumikan keberadaan dari fitur tertentu terhadap suatu kelas yang tidak terkait dengan fitur lainnya.

Persamaan Teorema Bayes:

$$P(C|X) = \frac{P(X|C).P(C)}{P(X)} \quad (2.1)$$

Keterangan:

P(C|X) : Probabilitas hipotesis berdasar kondisi (posteriori probability)

P(C) : Probabilitas hipotesis (prior probability)

P(X|C) : Probabilitas berdasarkan kondisi pada hipotesis

P(X) : Probabilitas X [19] .

Pada saat proses pengklasifikasian dokumen teks, maka pendekatan Bayes akan memilih kategori yang memiliki probabilitas paling tinggi (C MAP). Nilai p(x) dapat diabaikan karena nilainya adalah konstan untuk semua c. Dengan pendekatan Naive Bayes yang mengasumsikan bahwa setiap kata dalam setiap kategori adalah tidak tergantung satu sama lain, maka perhitungan dapat lebih disederhanakan. Berdasarkan hal tersebut maka rumus dapat disederhanakan menjadi persamaan 2.2.

$$C_{map} = \arg\max P_{(c)}. \prod_{i=1}^{n} P(W_j|C_j) \quad (2.2)$$

P(c) = Probabilitas prior.

P($W_j|C_j$) = Probabilitas kata pada kategori

Menghitung $P(c)$ diperlukan rumus pada persamaan 2.3.

$$P(c) = \frac{docsj}{Docs} \quad (2.3)$$

Docs j = jumlah dokumen pada kategori.

Docs = jumlah seluruh dokumen

Menghitung P($W_j|C_j$) diperlukan rumus pada persamaan 2.4.

$$P(W_j|C_j) = \frac{1+ni}{|c|+n(kosakata)} \quad (2.4)$$

ni = frekuensi kemunculan kata pada kategori.

C = jumlah kata dalam kategori

n(kosakata) = jumlah kata unik yang ada di dokumen [19].

2.1.11 TF-IDF

TF-IDF merupakan perhitungan yang terbagi menjadi dua yaitu *Term Frequency* dan *Inverse Document Frequency*. TF atau *Term Frequency* adalah sebuah *frequency* dari kata dalam sebuah dokumen yang semakin besar kemunculan term maka akan membuat bobot dan nilai kesesuaian semakin besar pula. Distribusi frasa dalam dokumen dihitung menggunakan algoritme yang disebut IDF, atau *Inverse Document Frequency*. Semakin kecil *term* pada dokumen semakin besar nilai IDF[20].

TF di hitung dengan melihat *frequency* kata dalam sebuah dokumen. Semakin besar kemunculan kata semakin besar pula bobot yang dihasilkan.

Rumus *Term Frequency* didefinisikan dengan:

$$TF(tk, dj) = f(tk, dj) \quad (2,5)$$

$f(tk, dj)$ = jumlah kemunculan term K pada dokumen j.

IDF digunakan untuk menghitung distribusi frasa, jumlah dokumen akan dibagi dengan jumlah dokumen yang mengandung term.

Rumus *Inverse Document Frequency* :

$$IDF(tk) = \log \frac{D}{\mathrm{df(t)}} \quad (2,6)$$

D = jumlah dokumen dataset

Df(t) = jumlah dokumen yang mengandung term.

TF-IDF menghitung hasil dari TF dikali dengan IDF.

Rumus TF-IDF:

$$TF\ IDF(tk, dj) = TF(tk, dj) * IDF(tk) \quad (2,7)\ [21]$$

2.1.12 K-Fold Cross Validation

K-Fold Cross Validation adalah metode yang digunakan untuk mengevaluasi kinerja dari sebuah algoritma dengan membagi data sampel secara acak dan mengelompokkannya sebanyak nilai k pada setiap *K-Fold*. Nilai k pada *K-Fold* adalah besar angka pemilah data untuk pembagi antara data *test* dan *train*[22].

2.1.13 Confusion Matrix

Confusion matrix adalah tahap evaluasi yang digunakan untuk melihat hasil akurasi atau performa dari klasifikasi. Langkah yang selalu digunakan adalah menilai *accuracy, precision, recall* dan *f1 score.*

Tabel 2.2 *Confusion Matrix*

Class	*Positive*	*Negative*
Positive	*True Positive* (TP)	*False Negative* (FN)
Negative	*False Positive* (FP)	*True Negative* (TN)

Accuracy adalah tingkat kedekatan antara nilai prediksi yang dihasilkan dengan nilai pada data sebenarnya [20].

Rumus *Accuracy*:

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TOTAL}} 100\% \quad (2,8)$$

Precision adalah tingkat ketepatan antara data yang diambil dengan informasi yang dibutuhkan [20].

Rumus *Precision*:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \quad (2,9)$$

Recall adalah tingkat keberhasilan sistem dalam menemukan kembali informasi [20].

Rumus *Recall*:

$$Recall = \frac{TP}{TP+FN} \quad (2,10)$$

f1 score adalah perbandingan rata-rata presisi dan recall yang dibobotkan [20].

Rumus *f1 score* :

$$f1\ score = 2x\frac{Precision \times Recall}{Precision + Recall} \quad (2,11)\ [23]$$