BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Beberapa penelitian terkait membahas hasil penelitian dengan topik serupa yaitu analisis sentimen ulasan Aplikasi Signal Samsat Pada Google Play Store Menggunakan Metode *Support Vector Machine*. Beberapa penelitian terdahulu yang terkait dengan topik penelitian ditinjau untuk mendukung penelitian ini. Berikut penelitian terdahulu yang berkaitan dan menjadi fokus dalam penelitian ini tercantum dalam tabel 1. Penelitian Terdahulu:

Table 1. Penelitian Terdahulu

No	Judul	Metode	Hasil
1.	Sentimen Analisis Review Aplikasi Digital Korlantas Pada Google Play Store Menggunakan Metode SVM [3].	*Support Vector Machine* (SVM)	Penelitian ini menggunakan data yang didapat melalui website Google Play dengan jumlah sampel 1200 data ulasan. Hasil yang didapatkan dengan menggunakan metode *Support Vector Machine* dengan rasio data 90:10 mendapatkan nilai akurasi 0.82, rasio 80:20 dan 60:40 mendapatkan nilai *accuracy* sama yaitu 0.74.
2.	Analisis Sentimen Ulasan Aplikasi Mola Pada Google Play Store Menggunakan Algoritma Support Vector Machine [7].	*Support Vector Machine* (SVM)	Penelitian ini menggunakan data yang dikumpulkan dari situs Google Play Store pada tanggal 1 Desember 2021 sampai 31 Januari 2022. Hasil yang diperoleh pada scenario 90:10 menggunakan kernel RBF (*Radial Basis Function)* menghasilkan *accuracy* 92,31%,

No	Judul	Metode	Hasil
			precision 96,3%, *recall* 89,66%¸dan *f1-score* 92,86%.
3.	Sentiment Analysis on Satusehat Application Using Support Vector Machine Method [8].	*Support Vector Machine* (SVM)	Penelitian ini menggunakan data yang dikumpulkan dari Google Play Store dengan jumlah 25000 data. Menghasilkan *accuracy* 0.91.
4.	Sentiment analysis for customer review: Case study of Traveloka [9]	*Support Vector Machine* (SVM), *Logistic Regression, Naïve Bayes* (NB)	Penelitian ini menggunakan data yang diperoleh dari sosial media twitter dengan menggunakan Twitter API. Dengan jumlah 1200 data diklasifikasi menjadi 610 data positif, dan 590 data negatif. Metode yang digunakan ialah *Support Vector Machine* (SVM), *Naïve Bayes* (NB), *Logistic Regression.* Berdasarkan hasil evaluasi *confusion matrix* dari ketiga metode tersebut, yang mendapatkan *accuracy* tertinggi adalah *Support Vector Machine.*
5.	Komparasi Payment Digital Untuk Analisis Sentimen Berdasarkan Ulasan Di Google Playstore Menggunakan Metode Support Vector Machine [10].	*Support Vector Machine* (SVM)	Penelitian ini menggunakan data yang diambil melalui API Google Play Scraper, dengan menggunakan metode *Support Vector Machine.* Data berjumlah 12000 dari dua aplikasi berbeda, masing-masing aplikasi berjumlah 6000 data. Pada aplikasi Dana memiliki hasil

No	Judul	Metode	Hasil
			accuracy 92%, sedangkan aplikasi Ovo hasil *accuracy* sebesar 90%.
6.	Penerapan Metode Naïve Bayes Classifier Dan Support Vector Machine Pada Analisis Sentimen Terhadap Dampak Virus Corona Di Twitter [11].	*Naïve Bayes Classifier* (NBC), *Support Vector Machine* (SVM)	Penelitian ini menggunakan data yang didapat menggunakan aplikasi *Rapidminer* dengan data seluruh pendapat Masyarakat mengenai dampak corona pada kehidupan sosial Masyarakat. Rentang pengumpulan data dari Januari – April 2021. Hasil yang diperoleh untuk algoritma *Naïve Bayes Classifier* menghasilkan akurasi sebesar 81.07%, dan untuk algoritma *Support Vector Machine* menghasilkan akurasi sebesar 79.96%.
7.	Sentiment Analysis Of Post-Covid-19 Inflation Based On Twitter Using The K-Nearest Neighbor And Support Vector Machine Classification Methods [12].	*K-Nearest Neighbor* (K-NN), *Support Vector Machine* (SVM)	Penelitian ini menggunakan data yang didapat dari twitter dengan melakukan proses *scraping* data *tweets* dengan berdasarkan kata kunci "Bahan Pokok Naik Pasca Pandemi", "BBM Naik", "Inflasi 2022", "Inflasi Covid19", dan "Inflasi Pasca Pandemi" yang didapat dari bulan Agustus – Oktober 2022. Terkumpul 5989 data dengan 2508 data positif, 1804 data negatif, dan 1677 data netral. Hasil yang didapat untuk metode *K-Nearest* Neighbor mendapat akurasi

No	Judul	Metode	Hasil
			sebesar 54%, dan untuk *Support Vector Machine* mendapat akurasi sebesar 79%.
8.	Sentiment Analysis Of Tourist Reviews Using K-Nearest Neighbors Algorithm And Support Vector Machine [13].	*K-Nearest Neighbors* (K-NN), *Support Vector Machine* (SVM)	Penelitian ini menggunakan data yang diperoleh dari hasil *scraping* dari situs Trip Advisor dengan lima (5) tempat wisata. Diperoleh hasil akurasi menggunakan metode *Support Vecrtor Machine* pada Waterbom Bali sebesar 88%, Mandala Suci Wenara Wana sebesar 83%, Tegalalang Terraces sebesar 89%, Pura Tanah Lot sebesar 88%, dan Pura Uluwatu sebesar 95%. Sedangkan dengan menggunakan metode *K-Nearest Neighbors* pada Waterbom Bali sebesar 67%, Mandala Suci Wenara Wana sebesar 60%, Tegalalang Terraces sebesar 69%, Pura Tanah Lot sebesar 71%, dan Pura Uluwatu sebesar 81%.
9.	Analisis Sentiment Aplikasi Gojek Menggunakan Support Vector Machine Dan K Nearest Neighbor [14].	*Support Vector Machine* (SVM), *K-Nearest Neighbor* (K-NN)	Data yang diperoleh pada penelitian ini berasal dari review pada aplikasi Gojek di website Google Playstore dilakukan menggunakan *web scraping* terdiri dari 2462 review. Hasil akurasi yang diperoleh menggunakan metode *Support Vector Machine* sebesar 87.98%,

No	Judul	Metode	Hasil
			dan metode *K-Nearest Neighbor* sebesar 82.14%.
10.	Analisis Sentimen Opini Terhadap Vaksin Covid-19 Pada Media Sosial Twitter Menggunakan Support Vector Machine Dan Naive Bayes [15].	*Support Vector Machine* (SVM), *Naïve Bayes* (NB)	Data yang diperoleh pada penelitian ini berasal dari data postingan twitter berdasarkan kata kunci vaksin Covid-19 yang diambil dari tahun 2021 dengan jumlah 1000 data. Hasil performa menggunakan algoritma *Support Vector Machine* memperoleh akurasi sebesar 90.47%, presisi 90.23%, dan recall 90.78%, sedamgkan algoritma *Naïve Bayes* memperoleh akurasi sebesar 88.64%, presisi 87.32%, dan recall 88.13%.

Berdasarkan tabel 1 penelitian terdahulu yang menjadi rujukan penelitian ini menunjukkan penggunaan metode yang sama dengan data dan hasil penngujian yang berbeda.

2.2 Data Mining

Data mining adalah proses yang menggunakan berbagai Teknik seperti Teknik statistik, matematika, kecedrasan buatan, dan machine learning untuk mengidentifikasi informasi pengetahuan yang berpotensi digunakan yang disimpan dalam database besar [16]. Data Mining digunakan untuk mengambil keputusan bisnis yang sangat penting dan bertujuan untuk menemukan pola yang tidak diketahui dalam data [17].

2.3 Text Mining

Text mining merupakan teknik yang digunakan dalam penyelesaian masalah klasifikasi. Teknik mining adalah satu proses analisis teks secara otomatis oleh komputer untuk mengekstrak informasi yang dirangkum dalam dokumen [18].

Secara umum, proses kerja text mining banyak mengadopsi metode dari penelitian data mining. Namun, perbedaannya adalah pola yang digunakan dalam text mining diperoleh dari Kumpulan teks dalam bahasa alami yang tidak terstruktur. Tahapan algoritma text mining dimulai dari praproses teks [19].

2.4 Analisis Sentimen

Analisis sentimen adalah pembelajaran komputer mengenai opini, perasaan, dan emosi yang diungkapkan dalam bentuk tekstual. Bertujuan untuk mengetahui apakah komentar dalam sekumpulan dokumen teks mengandung suatu objek yang bersifat positif atau negatif [20]. Analisis sentimen dalam Bahasa Indonesia merupakan suatu metode yang digunakan untuk mengidentifikasi dan memahami bagaimana suatu sentimen diekspresikan melalui teks. Selain itu, metode ini juga bertujuan untuk mengategorikan sentimen tersebut ke dalam dua jenis, yaitu positif dan sentimen negatif[21].

2.5 Google Play Store

Google Play Store adalah sebuah layanan konten digital, dimana pengguna smartphone yang memiliki sistem operasi android dapat mengunduh aplikasi maupun produk online lainnya secara gratis maupun berbayar [22].

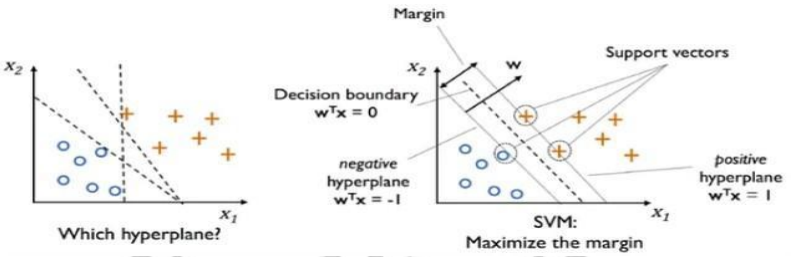
2.6 Scraping

Scraping adalah teknik mengambil data dan informasi dari internet yang dilakukan dengan bahasa pemrograman. Scraping umumnya berupa halaman situs website, dan menganalisis untuk diambil data tertentu dari halaman tersebut, kemudian disimpan dengan format .xlsx [23].

2.7 Support Vector Machine

Support Vector Machine (SVM) adalah salah satu model yang mengidentifikasi sampel data yang digunakan untuk klasifikasi. *Support Vector Machine* (SVM) memiliki konsep kerja dengan mencari hyperplane terbaik dengan memaksimalkan jarak antar kelas [6]. Hyperplane adalah suatu garis yang memisahkan data antara kelas atau kategori, dan margin adalah jarak antara hyperplane dan data terdekat dari masing-masing kelas. Sedangkan *Support Vector* adalah data yang paling dekat dengan hyperplane[24]. *Support Vector Machine* pada dasarnya digunakan dalam permasalahan klasifiaksi dua kelas atau

binary classification. Namun, diusulkan beberapa metode agar SVM dapat digunakan untuk klasifikasi berbagai kelas dengan menggabungkan beberapa *binary classifier*[25]. Konsep kerja dari algoritma *Support Vector Machine* dapat dilihat pada gambar 1.



Gambar 1. Konsep Kerja *Support Vector Machine*

Sumber : www.dsworld.org[26]

Metode *Support Vector Machine* secara efisien dapat digunakan untuk mengatasi data dengan karakteristik non-linier. Fungsi kernel berfungsi mentransformasikan data ke dimensi yang lebih tinggi, bertujuan meningkatkan struktur data dan memudahkan proses pemisahan. Rumus umum yang digunakan untuk SVM linear terdapat pada persamaan 1 berikut:

$$f(x) = sign\ (w.x + b) \qquad (1)$$

Pada persamaan ini, f(x) adalah fungsi prediksi, w merupakan vector normal pada *hyperplane*, x adalah vector fitur input, dan b adalah nilai bias atau *intercept* [27]. Permasalahan non-linier dapat diselesaikan dengan menerapkan trik kernel pada SVM, yang memungkinkan pemisahan kelas atau *hyperplane* menjadi dua kelas di ruang vector. Tabel dibawah merupakan persamaan dari masing-masing kernel [28].

Table 2. Persamaan Kernel SVM

Jenis Kernel	Model
Linear	$K(x.x') = x.x'$
Polynomial	$K(x.x') = (x.x' + c)'$
RBF	$K(x.x') = \exp(-y\|\|x - x'\|\|^2)$
Sigmoid	$K(x.x') = \tanh(ax.x' + \beta)$

2.8 TF-IDF

Pembobotan TF – IDF (*Term Frequency – Inverse Document Frequency*) digunakan untuk mengekstraksi dan memilih fitur. TF (*Term Frequency*) mengukur seberapa sering suatu kata muncul dalam suatu dokumen. Sedangkan IDF *(Inverse Document Frequency*) untuk menghitung seberapa penting suatu kata[29]. Dalam

proses perhitungan bobot menggunakan TF – IDF, nilai TF untuk setiap kata dihitung terlebih dahulu dengan bobot awal masing – masing kata sebesar 1. Sementara itu, nilai IDF dirumuskan sesuai dengan persamaan (2).

$$IDF(word) = \log \frac{td}{df} \qquad (2)$$

IDF(*word*) merupakan nilai IDF untuk setiap kata yang akan dihitung, di mana *td* adalah total jumlah dokumen yang tersedia, dan *df* adalah jumlah kemunculan kata tersebut di semua dokumen[30].

2.9 SMOTE

SMOTE merupakan metode yang sangat popular untuk mengatasi data yang tidak seimbang atau *imbalance* data dalam sebuah kelas. Metode ini kemudian menyeimbangkan dataset dengan menciptakan contoh baru dari kelas yang minoritas untuk meningkatkan kinerja metode klasifikasi[31]. Metode SMOTE digunakan untuk meningkatkan data, dan persamaan yang diterapkan adalah sebagai berikut ini [32]:

$$X_{syn} = X_i + (X_{knn} - X_i) * \sigma \quad (3)$$

Keterangan :

X_{syn} : data sintetis yang akan dibuat

X_i : data yang akan digandakan

X_{knn} : data dengan jarak terdekat dari data yang akan digandakan

σ: nilai acak antara 0 dan 1

2.10 Confusion Matrix

Dalam *machine learning*, *confusion matrix* adalah alat evaluasi visual berbentuk matriks yang berisi nilai prediksi yang benar dan nilai prediksi yang salah. Kolom pada *confusion matrix* berisi hasil kelas prediksi dan baris berisi hasil kelas yang sebenarnya[33]. Dalam pengujian untuk memperoleh hasil yang akurat, dalam *confusion matrix* terdapat beberapa nilai yang akan dievaluasi, antara lain presisi, *recall, dan f1 score*. Akurasi merupakan rasio antara jumlah prediksi tepat dengan total data yang dianalisa. Presisi merupakan rasio nilai prediksi yang benar – benar positif dibandingkan dengan semua hasil yang diprediksi positif. *Recall*

adalah perhitungan rasio prediksi benar positif terhadap semua data yang sebenarnya positif. Sedangkan *f1 score* adalah ukuran yang menggabungkan presisi dan *recall* [34].