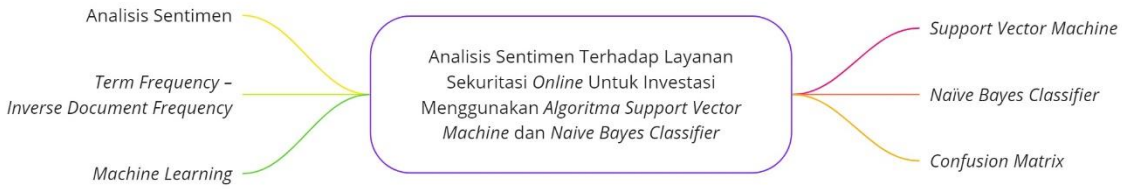
BAB II

TINJAUAN PUSTAKA

2.1 Landasan Teori

Penelitian itu berawal di masalah berakhir di pemecahan masalah (Wahono, 2021). Metode penelitian yang digunakan pada penelitian ini adalah metode penelitian kuantitatif. Maka perlu melakukan kajian teoritis dan studi pendahuluan terlebih dahulu pada objek untuk dapat memilih variabel apa yang akan diteliti. (Prof. Dr. Sugiyono, 2020). Berikut adalah *literature map* sebagai bahan kajian pustaka dalam penelitian ini.



Gambar 2.1 Peta Literasi Penelitian (Stefani. 2023)

2.1.1 Analisis Sentimen (*Sentiment Analysis*)

Analisis sentimen adalah bidang penelitian yang terdapat dalam pengolahan bahasa natural, komputasi lingustik dan text mining. Analisis sentimen atau opinion mining adalah studi komputasional dari opini orang lain, *appraisal*, serta emosi terdapat dalam entitas, *event* dan atribut yang dimiliki. Entitas yang dimaksud bisa berupa produk, individu, layanan, organisasi, kejadian, fenomena atau isu (Aggarwal, 2018). Analisis sentimen yang digunakan pada penelitian ini adalah

mengelompokkan polaritas pada suatu teks yang terdapat dalam sebuah dokumen atau kalimat kemudian dikemukakan bersifat positif, negatif atau netral.

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

Term Frequency–Inverse Document Frequency (TF-IDF) merupakan statistik numerik untuk mencerminkan pentingnya sebuah kata pada dokumen dalam sebuah kumpulan dokumen atau corpus (Rajaraman & Ullman, 2011). TF-IDF terdiri dari *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF).

Term Frequency (TF) merupakan jumlah frekuensi kemunculan kata dalam sebuah dokumen, dimana jumlah suatu kata yang muncul pada dokumen dibagi dengan jumlah kata yang terdapat pada dokumen, yang dirumuskan sebagai:

$$TF_{t,d} = \frac{f_{t,d}}{jumlah\ kata\ pada\ d} \qquad (2.1)$$

Dimana $f_{t,j}$ merupakan jumlah kata t yang muncul pada dokumen d.

Inverse Document Frequency (IDF) merupakan ukuran berapa banyak informasi yang diberikan oleh sebuah kata. Berbeda dengan *IDF* tradisional, *IDF sklearn* menggunakan konstanta kesatuan dalam penyebut dan pembilang. (Pedregosa et al., 2011a) yang dirumuskan sebagai:

$$idf(t) = \log\frac{1+n}{1+df(t)} + 1 \qquad (2.2)$$

Dimana n merupakan jumlah dokumen dalam sebuah korpus dan $df(t)$ merupakan jumlah munculnya kata i pada seluruh dokumen, dirumuskan sebagai:

$$TF.IDF_{(i,j,k)} = TF_{ij} \times IDF_i \qquad (2.3)$$

2.1.3 Machine Learning (ML)

Machine Learning adalah bagian dari *Artificial Intelegence* (AI) yang memungkinkan sebuah sistem untuk belajar dari data, bukan melalui pemrograman eksplisit (Hurwitz and Kirsch 2018). *Machine Learning* merupakan sebuah sistem yang dapat belajar dengan sendirinya. Sistem tersebut dapat memutuskan sesuatu dengan sendirinya tanpa harus ada campur tangan manusia. Hal ini menyebabkan komputer menjadi semakin pintar karena dapat belajar sendiri dari data yang dimilikinya. Dalam *machine learning*, data berperan sebagai bahan *input* untuk belajar (*training*) mengenai sesuatu untuk menghasilkan analisis yang benar. Data pada *machine learning* biasanya terbagi menjadi dua bagian, yaitu *data training* dan *data testing*. *Data training* digunakan untuk melatih algoritma *machine learning* sedangkan *data testing* digunakan untuk mengetahui performa dari algoritma yang digunakan. Dalam beberapa kasus terdapat *data validation* yang berperan sebagai bahan evaluasi untuk algoritma apabila hasil kurang maksimal.

Teknik *machine learning* dibagi menjadi beberapa macam, yaitu *supervised learning, unsupervised learning, reinforcement learning,* serta *neural network* dan *deep learning* (Hurwitz and Kirsch 2018). Pada *supervised learning,* data memiliki fitur berlabel yang mendefinisikan arti data. Sedangkan pada *unsupervised learning,* data tidak memiliki label. Teknik *unsupervised learning* memungkinkan mengklasifikasikan data berdasarkan klaster yang ditemukan. *Reinforcement learning* menerima umpan balik dari analisis data sehingga pengguna dipandu ke hasil terbaik. Teknik ini belajar melalui *trial* dan *error*

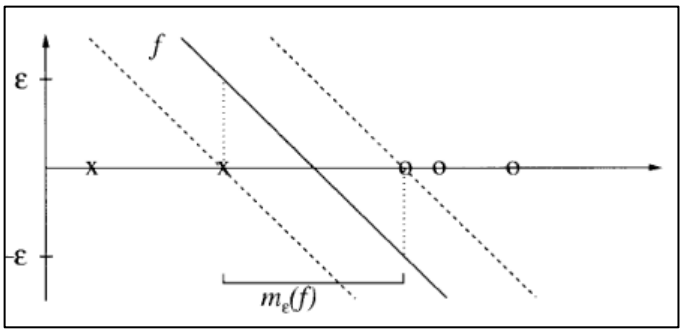
sehingga keputusan yang berhasil akan menghasilkan proses yang diperkuat. *Deep learning* adalah metode *machine learning* khusus yang menggabungkan jaringan saraf dalam lapisan yang berurutan untuk belajar dari data secara berulang yang memungkinkan menyelesaikan permasalahan dari data yang tidak terstruktur.

2.1.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah suatu teknik untuk melakukan sebuah prediksi, baik klasifikasi maupun regresi yang terinspirasi dari teori pembelajaran statistik (Vapnik, 1999). SVM membangun hyper-plane atau set hyper-plane dalam ruang dimensi tinggi atau tak terbatas, yang dapat digunakan untuk klasifikasi, regresi atau tugas lainnya (Pedregosa et al., 2011b) . Secara intuitif, pemisahan yang baik dicapai oleh hyper-plane yang memiliki jarak terbesar ke titik data pelatihan terdekat dari kelas mana pun (disebut margin fungsional), karena secara umum semakin besar margin, semakin rendah kesalahan generalisasi pengklasifikasi (Pedregosa et al., 2011b).

Banyak teknik *data mining* atau *machine learning* yang dikembangkan dengan asumsi kelinieran, sehingga algoritma yang dihasilkan terbatas untuk kasus-kasus yang linier. *Support Vector Machine* dapat bekerja pada data *non-linier* dengan menggunakan pendekatan kernel pada fitur data awal himpunan data. Berikut merupakan jenis-jenis fungsi kernel (Aggarwal, 2018).

1. Kernel Linear



Gambar 2.2 *Hyper-Plane* pada SVM (Schölkopf et al. 2000)

hyperplane klasifikasi linier SVM dinotasikan:

$$f(x) = w^T x + b \qquad (2.4)$$

Selain itu menurut (Vapnik dan Cortes, 1995) diperoleh persamaan:

$$[(w^T . x_i) + b] \geq 1 \; untuk \; y_i = +1$$
$$[(w^T . x_i) + b] \geq -1 \; untuk \; y_i = -1 \qquad (2.5)$$

dengan:

x_i = himpunan data training

i = 1, 2, …,n

y_i = label dari kelas x_i

2. Kernel Polinomial, digunakan untuk menyelesaikan masalah klasifikasi dimana dataset pelatihan sudah normal

3. Kernel Fungsi Gaussian Radial Basis (GRB), merupakan kernel yang paling banyak digunakan untuk menyelesaikan masalah klasifikasi untuk dataset yang tidak terpisah secara linier, dikarenakan akurasi pelatihan dan akurasi prediksi yang sangat baik pada kernel ini
4. Kernel Sigmoid, merupakan kernel trik SVM yang merupakan pengembangan dari jaringan saraf tiruan

2.1.5 Naïve Bayes Classifier (NBC)

Metode Naïve Bayes adalah metode klasifikasi dalam penambangan teks yang digunakan dalam analisis sentimen. Metode ini berpotensi baik dalam klasifikasi dalam hal presisi dan komputasi data. Naïve Bayes banyak digunakan dalam teknik klasifikasi adalah Unigram Naïve Bayes, Multinomial Naïve Bayes, dan Maximum Entropy Classification (Aggarwal, 2018). Berikut teorema umum naïve bayes yang digunakan dalam penelitian ini:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \qquad (2.6)$$

Keterangan:

X = data dengan *class* yang belum diketahui

H = hipotesis data merupakan suatu *class* spesifik

$P(H|X)$ = probabilitas hipotesis H berdasar kondisi X (posteriori probabilitas)

$P(H)$ = probabilitas hipotesis H (prior probabilitas)

$P(X|H)$ = probabilitas X berdasarkan kondisi pada hipotesis H

$P(X)$ = probabilitas X

Sedangkan distribusi yang digunakan adalah pengklasifikasi Guassian Naive Bayes. Istilah Multinomial Naive Bayes memberi tahu bahwa setiap P(Fi|C) adalah distribusi multinomial, bukan distribusi lainnya. Ini berfungsi dengan baik untuk data yang dapat dengan mudah diubah menjadi jumlah, seperti jumlah kata dalam teks. Ringkasnya, pengklasifikasi Naive Bayes adalah istilah umum yang mengacu pada independensi bersyarat dari setiap fitur dalam model, sedangkan pengklasifikasi Multinomial Naive Bayes adalah contoh spesifik dari pengklasifikasi Naive Bayes yang menggunakan distribusi multinomial untuk setiap fitur (Stuart J. Russell, 2021).

2.1.6 Confusion Matrix

Confusion matrix merupakan metode yang meringkas kinerja klasifikasi *clasifier* atau metode evaluasi sehubungan dengan beberapa data pengujian (Ting, 2010). Evaluasi performa model yang dibuat pada penelitian ini menggunakan metode *accuracy*, *precision*, *recall*, dan *F1-Score*. Untuk menghitung metode evaluasi yang sudah disebutkan, pada penelitian ini dibutuhkan *True Positive (TP), False Neutral1 (FNt1), False Negative1 (FNg1), False Positive1 (FP1), True Neutral (TNt), False Negative2 (FNg2), False Positive2 (FP2), False Neutral2 (FNt2), and True Negative (TNg*) dari *confusion matrix* setiap label (Yutika et al., 2021). Tabel 2.1 merupakan tabel matriks tiga dimensi yang terdiri dari tiga kelas, kelas positif, negatif, dan netral.

Tabel 2. 1 *Confusion Matrix* Tiga Kelas Sentimen

Actual	**Prediction**

	Positive	**Neutral**	**Negative**
Positive	True Positive (TP)	False Neutral1 (FNt1)	False Negative1 (FNg1)
Neutral	False Positive1 (FP1)	True Neutral (TNt)	False Negative2 (FNg2)
Negative	False Positive2 (FP2)	False Neutral2 (FNt2)	True Negative (TNg)

Sumber: (Saputro et al., 2018a)

Keterangan:

1. True positive : jumlah record positif yang diklasifikasikan sebagai positif,
2. True negative : jumlah record negatif yang diklasifikasikan sebagai negatif.
3. True netral : jumlah record netral yang diklasifikasikan sebagai netral.
4. False positive : jumlah record positif yang diklasifikasikan sebagai bukan positif,
5. False negative : jumlah record negatif yang diklasifikasikan sebagai bukan negatif.
6. False netral : jumlah record netral yang diklasifikasikan sebagai bukan netral.

Berdasarkan Tabel 2.1, penghitungan nilai *accuracy, precision, recall*, pada penelitian ini dapat dirumuskan sebagai berikut (Saputro et al., 2018b):

a. *Accuracy* (Akurasi)

   Accuracy adalah persentase jumlah *record* dari *train data* yang diklasifikasikan secara benar oleh algoritma, yang dinyatakan dalam persamaan 2.7.

$$Accuracy = \frac{TP+TNg+TNt}{TP+FNg1+FNt1+FP1+TNg+FNt2+FP2+FNg2+TNt} \; x\; 100\% \qquad (2.7)$$

b. *Recall*

Recall adalah evaluasi yang dilakukan untuk mengukur kelengkapan hasil klasifikasi, yang dinyatakan dalam

$$Recall.positive = \frac{TP}{TP+FNg1+FNt1} \; x\; 100\% \qquad (2.8)$$

$$Recall.negative = \frac{TNg}{FP1+TNg+FNt2} \; x\; 100\% \qquad (2.9)$$

$$Recall.neutral = \frac{TNt}{FP2+FNg2+TNt} \; x\; 100\% \qquad (2.10)$$

c. *Precision* (Presisi)

Precision adalah ukuran ketepatan hasil klasifikasi, yang dinyatakan dalam:

$$Precission.positive = = \frac{TP}{TP+FP1+FP2} \; x\; 100\% \qquad (2.11)$$

$$Precission.negative = = \frac{TNg}{FNg1+TNg+FNg2} \; x\; 100\% \qquad (2.12)$$

$$Precission.neutral = = \frac{TNt}{FNt1+FNt2+TNt} \; x\; 100\% \qquad (2.13)$$

d. *F-Measure atau f1 score*

Untuk menghitung *F1 – Score*, berdasarkan hasil penghitungan nilai *accuracy*, *precision*, *recall* dengan rumus di atas dapat dirumuskan sebagai berikut (Han et al., 2012):

$$F1 - Score.positive = 2\ x\ \frac{precision.Pst\ x\ recall.Pst}{precision.Pst + recall.Pst} \quad (2.14)$$

$$F1 - Score.negative = 2\ x\ \frac{precision.Ng\ x\ recall.Ng}{precision.Ng + recall.Ng} \quad (2.15)$$

$$F1 - Score.neutral = 2\ x\ \frac{precision.Nt\ x\ recall.Nt}{precision.Nt + recall.Nt} \quad (2.16)$$

2.2 *State of The Art* Bidang Penelitian

Tabel 2. 2 Matriks Penelitian

No	Nama Pengarang	Tahun	Judul	Isi Ringkasan	Hasil
1	(Samsir, Ambiyar, Unung Verawardina, Firman Edi, 2021)	2021	Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi Covid-19 Menggunakan Metode Naïve Bayes	Pembelajaran daring belum maksimal diterapkan di Indonesia pada masa pandemi yang terlihat dari tingginya kekecewaan public pada awal November 2020	Hasil penelitian ini menunjukkan bahwa pembelajaran daring memiliki 30% sentimen positif, 69% sentimen negatif dan 1% netral.
2	(Oryza Habibie Rahman et al., 2021)	2021	Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan SVM	Penggunaan kernel RBF menghasilkan nilai accuracy yang paling tinggi di antara kernel linear dan sigmoid.	Kernel RBF memiliki nilai accuracy 93%, nilai precision 84%, nilai recall sebesar 86%, dan nilai F-measure sebesar 83%.
3	Safitri (Juanita, 2020)	2020	Analisis Sentimen Persepsi Masyarakat	Algoritma yang digunakan adalah *Naïve Bayes* dengan menggunakan	Klasifikasi *Naïve Bayes* memiliki tingkat akurasi sebesar 82,90%.

No	Nama Pengarang	Tahun	Judul	Isi Ringkasan	Hasil
			Terhadap Pemilu 2019 Pada Media Sosial Twitter Menggunakan *Naïve Bayes*	perangkat lunak *Data Mining* WEKA untuk melakukan analisis sentimen persepsi masyarakat terhadap pemilu 2019	Diperoleh masing-masing sebesar 34,5% (471) tweet positif dan 65,5% (895) tweet negatif terhadap hasil *quick count*.
4	Rian Tineges, Agung Triayudi, Ira Diana Sholihati (Tineges et al., 2020)	2020	Analisis Sentimen Terhadap Layanan Indihome Berdasarkan Twitter Dengan Metode Klasifikasi *Support Vector Machine* (SVM)	Metode yang digunakan adalah *Support Vector Machine* dengan hasil yang dapat disimpulkan adalah tingkat kepuasan pengguna layanan Indihome cukup rendah.	Berdasarkan penelitian yang dilakukan diperoleh nilai accuracy 87%, precision 86%, recall 95%, error rate 13% dan F1-score 90%.
5	Yessi Yunita Sari, Aina Musdholifah, Anny Kartika Sari	2019	*Sarcasm Detection for Sentiment Analysis in Indonesian Tweets*	Penelitian ini menggunakan algoritma Random Forest. Ekstraksi fitur untuk analisis	Ada peningkatan rata-rata akurasi sebesar 5,49% dengan nilai akurasi sebasar 80,4%, presisi

No	Nama Pengarang	Tahun	Judul	Isi Ringkasan	Hasil
	(Yunitasari et al., 2019)			sentiment menggunakan TF-IDF dan klasifikasinya menggunakan algoritma Naïve Bayes.	sebesar 83,2% dan recall sebesar 91,3%.
6	Tati Mardiana, Hafiz Syahreva, Tuslaela (Mardiana et al., 2019b)	2019	Komparasi Metode Klasifikasi Pada Analisis Sentimen Usaha Waralaba Berdasarkan Data Twitter	Mengomparasi tingkat akurasi dalam mengklasifikasi opini masy dari data yang digunakan diambil di twitter yang berjumlah 1767 opini yang terdiri dari 1265 data positif dan 502 data negatif.	*Neural Network* dan *Support Vector Machine* menghasilkan akurasi tertinggi sebesar 83%. *Decision Tree* sebesar 81%, *Naïve Bayes* sebesar 80%, dan *K-Nearest Neighbors* sebesar 52%.
7	Styawati, Khabib Mustofa (Styawati & Mustofa, 2019)	2019	*A Support Vector Machine-Firefly Algorithm for Movie*	Metode yang digunakan dalam penelitian ini adalah SVM dengan menggunakan FA untuk mengoptimasi parameter SVM, dan	*Firefly Algorithm* dapat membantu SVM untuk mendapatkan kombinasi parameter yang sesuai

No	Nama Pengarang	Tahun	Judul	Isi Ringkasan	Hasil
			Opinion Data Classification.	metode Firefly sebagai metode optimasi parameter SVM.	berdasarkan akurasi dengan waktu eksekusi yang lebih singkat
8	Muh Amin Nurrohman, Azhari SN (Nurrohmat & SN, 2019)	2019	*Sentiment Analysis of Novel Review Using Long Short-Term Memory Method*	Mengklasifikasi review novel berbahasa Indonesia menggunakan metode LTSM yang dalam pengujiannya akan dibandingkan dengan metode Naïve Bayes.	Metode LTSM memiliki hasil yang lebih baik dibandingkan dengan *Naïve Bayes* dengan nilai akurasi 72,85%, presisi 73%, *recall* 72% dan *f-measure* 72%.
9	Mona Cindo, Dian Palupi Rini, Ermatita (Cindo et al., 2019)	2019	Studi Komparatif Metode Ekstraksi Fitur pada Analisis Sentimen Maskapai Penerbangan Menggunakan *Support*	Menganalisis sentiment dengan menambahkan lima fitur berbeda seperti topik pragmatic, lexical n-grams, POS, sentimen dan LDA, menggunakan dua metode	Pada metode *Maximum Entropy* menggunakan semua fitur ekstraksi dengan akurasi 92,7% dan pada *Support Vector Machine* akurasi yang diperoleh adalah 89,2%

No	Nama Pengarang	Tahun	Judul	Isi Ringkasan	Hasil
			Vector Machine dan *Maximum Entropy*	pembanding yaitu Support Vector Machine dan Maximum Entropy.	
10	Dinda Ayu Mutia (Ayu Muthia, 2018)	2018	Komparasi Algoritma Klasifikasi *Text Mining* Untuk Analisis Sentimen Pada Review Restoran	Menemukan informasi yang relevan dan tepat waktu dari berbagai review menggunakan algoritma Naïve Bayes dan Support Vector Machine dengan menambahkan fitur generate 2-grams (Bigrams) untuk itu dibuat aplikasi sederhana berbasis desktop menggunakan bahasa Java.	Algoritma *Naïve Bayes* lebih unggul dari algoritma *Support Vector Machine* dalam mengklasifikasi review restoran dengan teks berbahasa Indonesia. Akurasi algoritma *Naïve Bayes* mencapai 87%, algoritma *Support Vector Machine* menghasilkan akurasi sebesar 56%.
11	Tedy Agastya Dwi Permana, Firdaus	2017	Klasifikasi Emosi Teks Berbahasa Indonesia	*Maximum Entropy* menganalisa *query* dengan *dataset* pada	Dari uji coba sistem klasifikasi dengan menggunakan data *query*

No	Nama Pengarang	Tahun	Judul	Isi Ringkasan	Hasil
	Sholihin, Fika Hastarita (Tedy Agastya Dwi Permana, Firdaus Sholihin, 2017)		Menggunakan Metode *Maximum Entropy*	*database* untuk membentuk model probabilitas yang kemudian akan dilanjutkan dengan penentuan hasil klasifikasi. Proses uji coba pada sistem ini dilakukan dengan 3 cara yaitu uji coba sistem klasifikas *query*, *twitter* dan Data Sampel.	diperoleh hasil akurasi sebesar 93%, dan dengan menggunakan data *crawler* twitter diperoleh hasil akurasi sebesar 63%, kemudian dengan menggunakan data sampel diperoleh hasil rata-rata akurasi sebesar 64,6%.

2.3 Relevansi Penelitian

Tabel 2. 3 Tabel Relevansi Penelitian

Peneliti	**(Samsir, 2021)**	**(Oryza, 2021)**	**(Stefani, 2022)**
Judul	Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi Covid-19 Menggunakan *Metode Naïve Bayes*	Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan *Support Vector Machine*	Analisis Sentimen Terhadap Layanan Sekuritas Online untuk Investasi Menggunakan *Support Vector Machine* dan *Naïve Bayes Classifier.*
Masalah Penelitian	Sentimen terkait pembelajaran daring saat Covid-19 pada aplikasi *twitter*	Ujaran kebencian pada aplikasi twitter adalah perilaku yang dapat merugikan	Ulasan atau *review online* pada aplikasi Ajaib Sekuritas di *Google Play*.
Objek Penelitian	Klasifikasi *multi-class* pada media sosial *twitter* dan digunakan metode Naïve Bayes pada tahap klasifikasi sentimen dan interpretasi hasil analisis sentimen mengenai kasus Pembelajaran Daring saat Covid-19	Klasifikasi *multi-class* pada media sosial *twitter* setelah itu data masuk ke dalam proses pembobotan menggunakan TF-IDF dan klasifikasi menggunakan Support Vector Machine mengenai ujaran kebencian	Klasifikasi *multi-class* pada ulasan di aplikasi Ajaib Sekuritas dan dilakukan klasifikasi menggunakan algoritma *Support Vector Machine* dan *Naïve Bayes Classifier*

Algoritma / Metode	Menggunakan metode klasifikasi *Naïve Bayes*.	Menggunakan Support Vector Machine	Menggunakan metode *Naïve Bayes* dan *Support Vector Machine*
Implementasi	Klasifikasi *multi-class* dilakukan dengan bahasa pemrograman *python.*	Klasifikasi *multi-class* dilakukan dengan bahasa pemrograman *python.*	Klasifikasi *multi-class* dilakukan dengan bahasa pemrograman *python.*