BAB II

TINJAUAN PUSTAKA

2.1 State of The Art (SOTA)

Berdasarkan tinjauan terhadap studi-studi sebelumnya, jelas bahwa variasi dalam metode yang diterapkan dalam analisis sentimen akan mempengaruhi hasil akhir yang diperoleh. Setiap penelitian menggunakan metode berbeda yang mempengaruhi keakuratan dan efektivitas analisis sentimen. Untuk membandingkan berbagai metode secara lebih jelas yang digunakan, maka disusun tabel perbandingan berikut.

Tabel 2. 1 *State of The Art* (SOTA)

No	Author	Judul (Tahun)	Metode	Hasil
1.	Nurul Hijriani, Ermatita	[10]	*Naïve Bayes Classifier, Support Vector Machine*	Sebanyak 3.619 komentar yang dikumpulkan dari situs web Apify menunjukkan dominasi sentimen negatif (2.231 komentar), jauh melebihi sentimen positif yang tercatat sebanyak 1.437 komentar. Hasil akurasi yang diperoleh dari metode SVM sebesar 83.26% dan NBC sebesar 82.16%.
2.	Styawati, Nirwana Hendra, Auliya Rahman Isnain, Ari Yanti Ramadhani	[11]	*Support Vector Machine*	Data yang diuji pada penelitian sebanyak 302 data menggunakan pelabelan manual dan fitur TF-IDF. Hasil evaluasi menggunakan metode *Support Vector Machine* (SVM) yang dikombinasikan dengan kernel linear (98.67%) dan kernel RBF (98.34%).
3.	Sri Mulyani, Rice Novita	[12]	*Naïve Bayes Classifier*	Sebanyak 4.783 data yang telah melalui proses *labelling* manual dan ekstraksi fitur TF-IDF. Diuji dengan pembagian 3.826 data

No	Author	Judul (Tahun)	Metode	Hasil
				pelatihan dan 957 data uji. Pengujian ini menghasilkan akurasi klasifikasi Naïve Bayes sebesar 84,11%.
4.	Ritika Singh, Ayushka Tiwari	[13]	*Naïve Bayes Gaussian, Support Vector Machine, Logistic Regression, Decision Tree, K-Nearest Neighbour* (KNN), *Random Forest* (RF)	Sebanyak 1.500 data yang telah dipersiapkan dimasukkan ke dalam algoritma dengan menerapkan penimbangan *Count Vectorizer*. Hasil evaluasi dengan berbagai fitur yang dihasilkan seperti *unigram, bigrams*, dan *tri-grams*. Hasil yang dijabarkan merupakan akurasi terendah dan tertinggi dari fitur yang berbeda dan algoritma yang sama. 1) NBG: Terendah (79%) pada fitur B2. Tertinggi (87%) pada fitur A1, B1, dan C1. 2) SVM: Terendah (87%) pada fitur A1 dan B2. Tertinggi 88%) pada fitur lainnya. 3) LR: Terendah (84%) pada fitur A2. Tertinggi (88%) pada fitur B1 dan C1. 4) DT: Hasilnya sama pada semua fitur (87%). 5) KNN: Terendah (86%) pada fitur A2 dan B2. Tertinggi (87%) pada fitur lainnnya. 6) RF: Hasilnya sama pada semua fitur (88%).

No	Author	Judul (Tahun)	Metode	Hasil
5.	Nurul Rezki, Sri Astuti Thamrin, Siswanto	[14]	*Support Vector Machine* (SVM) dengan *Word2Vec*	Data penelitian terdiri dari 10.000 tweet yang diperoleh melalui proses *crawling* dari platform media sosial Twitter. Setelah itu, ekstraksi fitur dilakukan menggunakan *Word2Vec* untuk mendapatkan representasi vektor untuk kata dasar yang dihasilkan melalui proses tokenisasi. Hasil klasifikasi menunjukkan bahwa model SVM yang diimplementasikan dengan kernel RBF dan teknik *Word2Vec* menghasilkan akurasi sebesar 89.87%, presisi 91.20%, *recall* 84.44%, dan *F-measure* sebesar 87.68%.
6.	Asrikam Bulu, Elfira Umar, Paulus Miku Ate	[15]	*Naïve Bayes Classifier*	Dataset yang digunakan sebanyak 100 yang berasal dari hasil wawancara mahasiswa, namun setelah dilakukan proses *preprocessing* jumlah data berkurang menjadi 50 data dengan 31 positif dan 19 negatif. Hasil akurasi dari dataset dengan algoritma NBC sebanyak 82%, *precision* sebesar 72.73%, dan *recall* sebesar 84.21%.

Tabel 2.1 menampilkan algoritma *Naïve Bayes* yang populer dalam menganalisis sentimen terutama di berbagai bidang seperti opini publik dan pemrosesan bahasa alami. Selain itu ada algoritma lainnya seperti *Support Vector Machine, Logistic Regression*, dan lain-lainya sebagai pembanding dalam mengolah data teks. Meskipun, sejumlah studi sebelumnya telah berhasil membuktikan keefektifan berbagai metode dalam mengklasifikasikan teks ke kategori sentimen netral, positif, dan negatif. Sentimen publik terhadap inovasi kesehatan terutama kebijakan penyebaran nyamuk *Wolbachia* di Indonesia masih terbilang terbatas. Oleh sebab itu, tujuan dari penelitian ini untuk menganalisis dan menerapkan kinerja algoritma *Naïve Bayes* dalam konteks masalah ini. Diharapkan, penelitian ini dapat memberikan gambaran yang lebih luas tentang tingkat penerimaan publik terhadap upaya pengendalian

penyakit DBD yang semakin mendesak dan relevan di tengah lonjakan kasus demam berdarah di Indonesia.

2.2 Machine Learning

Bagian dari bidang kecerdasan buatan (AI) yang disebut *Machine Learning* berkonsentrasi pada pembuatan model dan algoritma untuk memungkinkan komputer untuk belajar dari data secara mandiri, sehingga mampu melakukan fungsi prediktif atau pengambilan keputusan tanpa memerlukan pemrograman eksplisit. Dengan memanfaatkan pola yang terdapat dalam data, algoritma machine learning dapat meningkatkan kinerjanya seiring dengan bertambahnya volume data yang diproses. Penggunaan utama machine learning adalah untuk melatih sistem agar dapat mengelola dan menganalisis data secara lebih efektif. Namun, tantangan signifikan dapat muncul dalam menafsirkan informasi yang diekstraksi dari data, terutama ketika model harus menangani dataset yang kompleks atau sangat ambigu [16].

2.3 Analisis Sentimen

Menganalisis sentimen atau juga dikenal penambangan opini merupakan subbidang penambangan teks yang berfokus pada penentuan persepsi publik atau subjektivitas terhadap suatu isu tertentu [17]. Mengklasifikasikan polaritas teks baik dari opini, kalimat, atau dokumen adalah tugas utama dalam analisis sentimen untuk menentukan teks tersebut mengandung makna netral, negatif, atau positif [18]. Tujuan utama disiplin ini adalah untuk memahami pola opini publik dan menganalisis respons orang terhadap topik, produk, atau peristiwa. Selain manfaat tersebut, analisis sentimen juga memiliki fungsi strategis dalam mendeteksi pergeseran tren opini, memantau citra merek, mengevaluasi efektivitas komunikasi, dan mengukur kesuksesan suatu kebijakan atau strategi pemasaran.

Dalam penerapannya, analisis ini tidak hanya bergantung pada kata-kata individual, namun juga pada konteks, nada, bahkan gaya bahasa yang digunakan termasuk identifikasi emosi yang lebih kompleks seperti sarkasme, satir, dan ironi.

2.4 Data *Preprocessing*

Preprocessing data merupakan fase awal yang krusial dalam suatu metode. Tujuannya adalah untuk membuat data mentah menjadi format yang lebih terstruktur sehingga memudahkan tahap analisis selanjutnya [17]. Kualitas pengolahan data pada tahap ini memiliki dampak signifikan terhadap kemampuan generalisasi algoritma pembelajaran mesin [18]. Oleh karena itu, tahap ini memastikan bahwa data yang digunakan berkualitas tinggi, konsisten, dan siap untuk analisis lebih lanjut yang pada akhirnya akan meningkatkan kinerja model dan akurasi analisis. Mengingat kualitas data sangat menentukan hasil akhir analisis atau pemodelan, tahap prapemrosesan ini dianggap sebagai tahap krusial dalam proses analisis data. Teknik yang umum digunakan dalam pemrosesan data meliputi pembersihan data, *case folding*, normalisasi, *stopword removal*, tokenisasi, dan *stemming*. Penjelasan lebih rinci mengenai teknik-teknik pemrosesan data yang diimplementasikan akan disajikan pada Tabel 2.2.

Tabel 2. 2 Penjelasan Teknik *Preprocessing*

Proses	**Penjelasan**
Cleaning	Penghilangan angka, tanda baca (delimiter), *uniform resource locator* (URL), dan tagar (#) dilakukan karena masih banyak pengguna yang menggunakan simbol-simbol tersebut dalam komentar [12].
Case Folding	Proses standarisasi yang mengganti semua karakter di dokumen menjadi huruf kecil, sementara karakter lain diperlakukan sebagai pemisah atau pembatas [11].
Normalization	Proses mengembalikan kata-kata non-standar seperti istilah informal (*slang*) atau singkatan kembali ke kata-kata standarnya [19].
Tokenization	Proses yang menerima masukan teks dan mengubahnya menjadi serangkaian token (kata) dengan memecah teks berdasarkan spasi atau tanda baca [20].
Stopword Removal	Proses yang menghilangkan kata-kata yang dianggap kurang penting guna meningkatkan kualitas analisis teks dengan menyoroti kata-kata yang memiliki makna atau relevansi yang lebih besar [21].
Stemming	Proses linguistik komputasional yang mengubah sebuah kata menjadi bentuk dasarnya (akar kata) dengan menghilangkan semua afiks, baik prefiks maupun sufiks [22].

2.5 Kappa Statistik

Kappa statistik atau *Cohen's Kappa* adalah metrik statistik guna mengukur tingkat kesepakatan (*agreement*) antara pengamat atau metode klasifikasi dengan menghitung kemungkinan kesepatan yang terjadi secara kebetulan. Validitas hasil pelabelan dinilai menggunakan *Cohen's Kappa* untuk mengukur keakuratan dan konsistensi antar penilai [21]. Perhitungan didasarkan pada perbandingan antara kesepakatan yang sebenarnya (*observed agreement*) dengan kesepakatan yang diharapkan (*expected agreement*).

Tabel 2. 3 Variasi Antar Penilai

		Hasil Penilai 1			
Hasil Penilai 2	Sentimen	Positif	Netral	Negatif	Total
	Positif	a	d	g	y_0
	Netral	b	e	h	y_1
	Negatif	c	f	i	y_2
	Total	x_0	x_1	x_2	z

Penjelasan tabel 2.3 diatas:

a, e, dan i : Jumlah dua pengamat setuju.
d, g, b, h, c, dan f : Jumlah dua pengamat tidak setuju
z : Total data yang diklasifikasikan

Apabila tidak terdapat perbedaan atau semua pihak setuju, maka nilai d, g, b, h, c, dan f adalah nol. Sehingga kesepakatan yang sebenarnya (*observed agreement*) yang dinotasikan dengan p_0 menjadi 1. Sebaliknya, jika tidak ada kesepakatan maka a, e, dan i adalah nol, serta kesepakatan yang sebenarnya (*observed agreement*) yang dilambangkan p_0 menjadi 0.

1. Rumus *Observed agreement* pada persamaan 2.5.1.

$$p_0 = \frac{a + e + i}{z} \qquad (2.5.1)$$

2. Rumus *Expected agreement* pada persamaan 2.5.2.

$$p_e = \left[\left(\frac{x_0}{z}\right) * \left(\frac{y_0}{z}\right)\right] + \left[\left(\frac{x_1}{z}\right) * \left(\frac{y_1}{z}\right)\right] + \left[\left(\frac{x_2}{z}\right) * \left(\frac{y_2}{z}\right)\right] \qquad (2.5.2)$$

3. Rumus *Kappa* pada persamaan 2.5.3.

$$K = \frac{(p_0 - p_e)}{1 - p_e} \qquad (2.5.3)$$

Hasil perhitungan *kappa* ditafsirkan menurut tabel 2.4 [23].

Tabel 2. 4 Interpretasi Kappa

NILAI KAPPA	KEERATAN KESEPAKATAN
< 0,2	Rendah (*Poor*)
0,21 – 0,40	Lumayan (*Fair*)
0,41 – 0,60	Cukup (*Moderate*)
0,61 – 0,80	Kuat (*Good*)
0,81 – 1,00	Sangat Kuat (*Very Good*)

Label akhir untuk setiap data dipilih berdasarkan *consensus vote* (kesepakatan bersama) dari dua penilai. *Consensus vote* adalah proses diskusi dan penyelarasan persepsi antara kedua penilai untuk mencapai keputusan bersama. Dalam metodologi ini, kedua penilai akan meninjau dan menganalisis konteks komentar yang memiliki label berbeda lalu berdiskusi untuk mencapai kesepakatan dan menentukan label akhir yang paling tepat sesuai dengan pedoman pelabelan yang ditetapkan.

2.6 TF-IDF (*Term Frequency-Inverse Document Frequency*)

Untuk mengetahui seberapa penting atau relevan kata dalam dokumen yang merupakan bagian dari keseluruhan korpus digunakan metode pembobotan statistik yang dikenal sebagai TF-IDF [24]. Prinsip intinya adalah bobot suatu kata akan meningkat jika sering muncul dalam dokumen terkait, tetapi akan menurun jika kata tersebut terlalu umum di seluruh korpus [25]. Dengan menggabungkan komponen TF dan IDF, TF-IDF secara efektif mengidentifikasi istilah yang paling signifikan dengan mengurangi penggunaan kata-kata yang biasa dan tidak informatif.

Seberapa sering sebuah kata muncul dalam dokumen tertentu diukur dengan istilah *Term Frequency*. Meskipun kemunculan kata yang sering akan menghasilkan nilai TF yang lebih tinggi, TF dihitung relatif terhadap panjang dokumen untuk mencegah pembobotan kata yang berulang secara berlebihan dalam dokumen yang panjang. Sementara itu, *Inverse Document Frequency* mengukur keunikan atau kelangkaan kata di seluruh dokumen dalam korpus. Semakin sedikit dokumen yang memuat sebuah kata, semakin tinggi nilai IDF-nya yang menunjukkan relevansi atau kepentingan yang lebih besar dalam konteks tertentu. Semua langkah ini dilakukan sebagai bagian dari metode TF-IDF untuk proses pembobotan kata.

a) Perhitungan TF menggunakan persamaan 2.6.1

$$TF(t_i, d) = \frac{frekuensi\ t\ pada\ d}{jumlah\ kata\ pada\ d} \qquad (2.6.1)$$

Keterangan:

- t_i : Kata ke-i
- d : Dokumen

b) Perhitungan IDF menggunakan persamaan 2.6.2

$$IDF(t_i) = \ln \frac{N}{DF(t_i)} \qquad (2.6.2)$$

Keterangan:

- t_i : Kata ke-i
- ln : Fungsi algoritma alami
- N : Jumlah total dokumen yang dikumpulkan
- DFt_i : Jumlah dokumen dalam koleksi yang t_i miliki

c) Perhitungan dengan perkalian dari TF dan IDF menggunakan persamaan 2.6.3

$$TFIDF = TF(t_i, d) \times IDF(t_i) \qquad (2.6.3)$$

2.7 *Multinomial Naive Bayes*

Untuk kasus klasifikasi teks, algoritma *Multinomial Naïve Bayes* berbasis pada prinsip probabilitas [26], dimana metode ini berfungsi untuk menghitung probabilitas suatu kelas dengan mengacu pada frekuensi kemunculan kata dalam suatu dokumen. Penggunaan frekuensi kemunculan kata sebagai fitur

dokumen dipresentasikan sebagai *vector* frekuensi kata-kata. Pengkategorian dokumen, analisis sentiment, dan filtrasi spam adalah beberapa contoh masalah klasifikasi yang melibatkan data berbentuk *bag of words*. Algoritma ini sangat cocok untuk tugas-tugas tersebut. Fitur-fitur *Multinomial Naïve Bayes* bersifat independen satu sama lain dan probabilitas mengikuti distribusi *multinomial*. Rumus proses perhitungan metode *naïve bayes multinomial* sebagai berikut.

a) Menghitung probabilitas suatu dokumen d menjadi bagian kelas c dengan persamaan 2.7.1

$$P(c|d) = \frac{P(c) \times P(d|c)}{P(d)} \qquad (2.7.1)$$

Keterangan:

- $P(c|d)$: Probabilitas posterior
- $P(c)$: Probabilitas prior
- $P(d|c)$: Kemungkinan dalam kelas c terdapat dokumen d
- $P(d)$: Kemungkinan dokumen d

b) Menghitung $P(d|c)$ dengan persamaan 2.7.2

$$P(d|c) = \prod_{i=1}^{n} P(t_i|c)^{f_i} \qquad (2.7.2)$$

Keterangan:

- t_i : Kata ke-i
- f_i : Frekuensi kemunculan kata t_i
- $P(t_i|c)$: Probabilitas kata t_i muncul dalam kelas c

c) Menghitung $P(t_i|c)$ dengan persamaan 2.7.3

$$P(t_i|c) = \frac{N_{t_i,c} + \alpha}{N_c + \alpha V} \qquad (2.7.3)$$

Keterangan:

- $N_{t_i,c}$: Jumlah kemunculan kata t_i dalam semua dokumen yang termasuk kelas c
- N_c : Total jumlah kata dalam semua dokumen yang termasuk kelas c
- V : Ukuran dari seluruh vocabulary (total kata unik)
- α : Parameter smoothing

2.8 Confusion Matrix

Tabel *confusion matrix* digunakan untuk mengevaluasi bagaimana model klasifikasi dalam pembelajaran mesin bekerja. Tabel ini menunjukkan perbandingan antara prediksi yang dihasilkan oleh model dan nilai data yang diuji secara aktual. Menurut [21], *confusion matrix* digunakan untuk mengukur evaluasi atau akurasi klasifikasi dengan cara memeriksa nilai prediksi setiap kelas, kemudian membandingkannya dengan kelas aslinya. Oleh karena itu, teknik ini dianggap krusial dalam memahami kekuatan dan keterbatasan suatu model. Dalam konteks penelitian ini, *confusion matrix* multikelas digunakan karena melibatkan pengklasifikasian data ke dalam tiga sentimen yaitu netral, positif, dan negatif seperti yang disajikan pada Tabel 2.5.

Tabel 2. 5 *Confusion Matrix Multiclass*

Prediksi	**Aktual**		
	Positif	**Netral**	**Negatif**
Positif	TPos	FPos	FPos
Netral	FNeu	TNeu	FNeu
Negatif	FNeg	FNeg	TNeg

Berdasarkan tabel 2.5 perhitungan evaluasi dengan empat output menggunakan rumus-rumus berikut.

1. Akurasi

   Akurasi didefinisikan sebagai proporsi total prediksi yang diklasifikasikan terhadap jumlah total data yang diuji dengan benar. Nilai akurasi ini dapat dihitung menggunakan persamaan 2.8.1.

$$\text{Akurasi} = \frac{TPos + TNeu + TNeg}{TPos + TNeu + TNeg + FPos + FNeu + FNeg} \quad (2.8.1)$$

   Keterangan:

   - TP (*True Positive*) : Sesuai prediksi positif
   - TNeu (*True Neutral*) : Sesuai prediksi netral
   - TN (*True Negative)* : Sesuai prediksi negatif

- FP (*False Positive*) : Tidak sesuai dengan prediksi positif (seharusnya bukan positif)
- FNeu (*False Neutral*) : Tidak sesuai dengan prediksi netral (seharusnya bukan netral)
- FN (*False Negative*) : Tidak sesuai dengan prediksi negatif (seharusnya bukan negatif).

2. Presisi (*precision*)

Presisi mengukur perbandingan antar dugaan positif yang benar dibandingkan dengan total prediksi positif dan metrik ini khususnya relevan untuk skenario dengan tingkat kesalahan tipe I yang tinggi (*false positive*). Presisi dihitung menggunakan persamaan 2.8.2.

$$\text{Presisi} = \frac{TPos}{TPos + FPos} \qquad (2.8.2)$$

3. *Recall* (*Sensitivity* atau *True Positive Rate*)

Recall juga dikenal sebagai sensitivitas atau rasio positif sejati berfungsi sebagai metrik yang mengukur proporsi kasus positif yang diidentifikasi secara akurat oleh model. Pengukuran ini krusial dalam situasi di mana rasio kesalahan tipe II (*false negative*) cenderung tinggi dan dapat dihitung menggunakan persamaan 2.8.3.

$$\text{Recall} = \frac{TPos}{TPos + FNeu + FNeg} \qquad (2.8.3)$$

4. F1-*Score*

F1-*score* adalah metrik kombinasi yang menyelaraskan presisi dan perolehan kembali menjadi satu skor tunggal menjadikannya alat evaluasi yang sangat berguna ketika menangani data yang tidak seimbang. persamaan 2.8.4 menunjukkan hasil F1-*score*.

$$\text{F1 Score} = 2 \times \frac{\text{Presisi} \times \text{Recall}}{\text{Presisi} + \text{Recall}} \qquad (2.8.4)$$