BAB II
TINJAUAN PUSTAKA

3.1 Kajian Pustaka

Dalam menyusun penelitian, tentu diperlukan pemahaman dan penggunaan teori-teori yang terkait dengan masalah dan ruang lingkup yang akan dibahas dalam penelitian. Salah satu cara yang dilakukan adalah dengan mengkaji penelitian-penelitian sebelumnya, yang bertujuan untuk mendapatkan pemahaman yang lebih dalam dan sebagai acuan dalam penyusunan penelitian ini. berikut merupakan referensi jurnal terdahulu berkaitan analisa sentimen menggunakan algoritma N*aive Bayes*.

Penelitian yang dilakukan oleh Elisa Febriyani dan Herny Februariyanti dengan judul "Analisis Sentimen Terhadap Program Kampus Merdeka Menggunakan Algoritma Naive Bayes Classifier Di Twitter" [15]. Penelitian tersebut bertujuan menganalisis sentimen opini publik terhadap program kampus merdeka di twitter untuk mengetahui tingkat akurasi pada metode serta persentase sentimen sebagai evaluasi pada algoritma, kinerja dan program kampus merdeka ini sendiri. Berdasarkan hasil penelitian, pengklasifikasian yang dapat dilakukan oleh sistem mendapatkan hasil klasifikasi sentimen positif sebanyak 272 opini dan sentimen negatif sebanyak 229 opini dengan rata - rata akurasi 60%, presisi 64%, *recall* 58% dan *f1 - score* 58%.

Selanjutnya penelitian yang dilakukan oleh Alfandi Safira dan Firman Noor Hasan dengan judul "Analisis Sentimen Masyarakat Terhadap Paylater Menggunakan Metode Naive Bayes Classifier" [12]. Dalam penelitiannya bertujuan mengetahui pandangan masyarakat terhadap *paylater*. Pengujian model dengan *confussion matrix* menunjukan bahwa algoritma *Naive Bayes Classifier* sebesar 91%.

Selanjutnya penelitian yang dilakukan oleh Murni, Imam Riadi, dan Abdul Fadlil dengan judul "Analisis Sentimen HateSpeech pada Pengguna Layanan Twitter dengan Metode Naive Bayes Classifier (NBC)" [16]. Tujuan dari penelitian adalah untuk melakukan analisis sentimen *HateSpeech* pengguna layanan twitter dengan metode *Naive Bayes Classifier* Berdasarkan evaluasi menggunakan *Confusion Matrix* diperoleh akurasi tertinggi sebesar 80%, *precision* 100%, *recall* 80%, dan *F1-Score* 89% pada skenario pengujian data *training* 70% dan *testing* 30%.

Selanjutnya penelitian yang dilakukan oleh Debora Chrisinta dan Justin Eduardo Simarmata yang berjudul "Analisis Sentimen Penilaian Masyarakat Terhadap Pejabat Publik Menggunakan Algoritma Naive Bayes Classifier" [17]. Tujuan penelitian ini dilakukan adalah menerapkan algoritma *Naive Bayes* dalam melakukan klasifikasi sentimen data Twitter penilaian masyarakat terhadap pejabat publik. Hasil analisis sentimen menunjukkan penilaian masyarakat dengan frekuensi tertinggi berada pada kelas negatif. Performa algoritma menunjukkan nilai *accuracy* sebesar 64,55% dengan *error rate* sebesar 35,45%.

Selanjutnya penelitian yang dilakukan oleh Alexandre Liberti Duarte Tavares dan Eddy Nurraharjo dengan judul "Analisis Sentimen Dan Klasifikasi Tweet Terkait Naiknya Kasus Omicron Menggunakan Naive Bayes Classifier" [18]. Tujuan penelitian ini dilakukan adalah menghitung tingkat akurasi opini pengguna twitter, dengan Metode klasifikasi *Naive Bayes*. Metode ini mengklasifikasikan data sentimen yang diambil dari data positif dan data negatif dari twitter yang diimput menggunakan twitter *API* dengan kata "OMICRON". Hasil klasifikasi data penelitian ini memiliki nilai dari klasifikasi data *tweet* menghasilkan nilai akurasi sebesar 90.0%, nilai *precision* 90.00%, nilai *recall* 90,00% dan nilai *f1-score* 90,00%.

Tabel 2.1 Penelitian Terdahulu

No	Judul	Tahun	Peneliti	Hasil
1.	Analisis Sentimen Terhadap Program Kampus Merdeka Menggunakan Algoritma Naive Bayes Classifier Di Twitter	2023	Elisa Febriyani dan Herny Februariyanti	Hasil penelitian, pengklasifikasian yang dapat dilakukan oleh sistem mendapatkan hasil klasifikasi sentimen positif sebanyak 272 opini dan sentimen negatif sebanyak 229 opini dengan rata - rata akurasi 60%, presisi 64%, *recall* 58% dan *f1 - score* 58%.
2.	Analisis Sentimen Masyarakat Terhadap Paylater Menggunakan Metode Naive Bayes Classifier	2023	Alfandi Safira dan Firman Noor Hasan	Pengujian model dengan *confussion matrix* menunjukan bahwa algoritma *Naive Bayes Classifier* sebesar 91%.
3.	Analisis Sentimen HateSpeech pada Pengguna Layanan Twitter dengan Metode Naive Bayes Classifier (NBC)	2023	Murni, Imam Riadi, dan Abdul Fadlil	Berdasarkan evaluasi menggunakan *Confusion Matrix* diperoleh akurasi tertinggi sebesar 80%, *precision* 100%, *recall* 80%, dan *F1-Score* 89% pada skenario pengujian data *training* 70% dan *testing* 30%.
4.	Analisis Sentimen Penilaian Masyarakat Terhadap Pejabat Publik Menggunakan Algoritma Naive Bayes Classifier	2023	Debora Chrisinta dan Justin Eduardo	Hasil analisis sentimen menunjukkan penilaian masyarakat dengan frekuensi tertinggi berada pada kelas negatif. Performa algoritma menunjukkan nilai *accuracy* sebesar 64,55% dengan *error rate* sebesar 35,45%.

No	Judul	Tahun	Peneliti	Hasil
5.	Analisis Sentimen Dan Klasifikasi Tweet Terkait Naiknya Kasus Omicron Menggunakan Naive Bayes Classifier	2023	Alexandre Liberti Duarte Tavares dan Eddy Nurraharjo	Hasil klasifikasi data penelitian ini memiliki nilai dari klasifikasi data *tweet* menghasilkan nilai akurasi sebesar 90.0%, nilai *precision* 90.00%, nilai *recall* 90,00% dan nilai *f1-score* 90,00%.
6.	Analisa Sentimen Pengguna Transportasi Jakarta Terhadap Transjakarta Menggunakan Metode Naives Bayes dan K-Nearest Neighbor	2023	Ismia Iwandini, Agung Triayudi dan Gatot Soepriyono	Hasil akurasi pendekatan Naive Bayes untuk analisis sentimen data Twitter terkait penggunaan transportasi Transjakarta sebesar 61,1%, dan akurasi metode K-Nearest Neighbor sebesar 75,7%. Hasil analisis sentimen yang mengungkapkan bahwa sikap positif mendominasi, semakin menunjukkan bahwa masyarakat umum peduli dengan adanya teknologi. Dalam penyelidikan ini, ditentukan bahwa pendekatan K-Nearest Neighbor lebih baik daripada Naives Bayes.

Berdasarkan Tabel 2.1 merupakan penelitian sebelumnya yang digunakan sebagai referensi untuk mengembangkan skema penulisan penelitian ini karena memiliki keterkaitan dengan topik penelitian.

3.2 Dasar Teori

2.2.1 Pemindahan Ibu Kota

Pada pidato kenegaraan tanggal 16 Agustus 2019, Presiden Republik Indonesia mengumumkan rencana pemindahan Ibu kota Republik Indonesia dan meminta izin kepada Majelis Permusyawaratan Rakyat. Rencana tersebut telah melalui kajian dari Badan Perencanaan dan Pembangunan Nasional (Bappenas RI). Presiden menekankan bahwa Ibu kota baru bukan hanya sebagai simbol identitas bangsa, tetapi juga sebagai perwujudan kemajuan bangsa. Dengan lokasi Ibu kota baru yang terletak di tengah Indonesia, diharapkan dapat mewujudkan pemerataan ekonomi dan pembangunan yang adil. Pada Senin, 26 Agustus 2019, Presiden Republik Indonesia dalam keterangannya memutuskan bahwa sebagian wilayah Penajam Paser Utara dan sebagian Kutai Kartanegara di Kalimantan Timur menjadi lokasi pembangunan ibu kota baru Republik Indonesia [19].

2.2.2 *Machine Learning*

Machine learning merupakan sub-bidang dalam keilmuan kecerdasan buatan (*Artificial Intelligence*) yang telah mendapatkan banyak perhatian dan penelitian dalam upaya memecahkan berbagai masalah. Ulasan dari berbagai bidang dihadirkan dalam bentuk pemecahan masalah dan algoritma yang digunakan, dan biasanya dikategorikan menjadi tiga jenis utama dalam *machine learning*, yaitu *supervised learning*, *unsupervised learning*, dan *reinforcement learning*. Namun, dalam ulasan ini fokus terbatas pada beberapa bidang tertentu, dan hasilnya menunjukkan bahwa bidang kedokteran atau medis merupakan bidang yang paling dominan dalam perkembangan terkini, meskipun terdapat juga beberapa bidang lain yang menjadi perhatian, seperti industri, teknologi, dan lalu lintas [20].

2.2.3 Analisis Sentimen

Analisis sentimen adalah cabang dari penambangan data yang bertujuan untuk menganalisis dan memproses data teks dalam bentuk pendapat atau ulasan tentang berbagai entitas seperti produk, layanan, organisasi, individu, dan topik tertentu. Dalam analisis ini, algoritma dan metode komputasional digunakan untuk mengklasifikasikan dan menginterpretasikan sentimen yang terkandung dalam data tekstual, seperti apakah pendapat tersebut bersifat positif, negatif, atau netral. Analisis sentimen memiliki berbagai aplikasi, mulai dari pemantauan citra merek, pengelolaan reputasi perusahaan, hingga pengembangan strategi pemasaran yang lebih efektif, dengan tujuan untuk mendapatkan wawasan berharga dalam pengambilan keputusan dan merespons umpan balik dari pengguna atau konsumen [21].

2.2.4 *Preprocessing*

Preprocessing merupakan tahapan yang sangat penting dalam proses pengolahan data. Data yang digunakan seringkali tidak dalam kondisi yang ideal untuk diproses secara langsung. Terkadang, data tersebut menghadapi berbagai masalah seperti *missing value, data redundant, outliers,* atau format data yang tidak sesuai dengan sistem. Semua permasalahan tersebut dapat mengganggu hasil dari proses pengolahan data itu sendiri. Oleh karena itu, untuk mengatasi masalah tersebut, diperlukan tahapan *preprocessing* yang bertujuan untuk membersihkan, menormalkan, dan mengubah data menjadi format yang lebih sesuai dan siap untuk diolah [22]. *Preprocessing* ini melibatkan serangkaian langkah yang meliputi *cleaning, case folding, tokenisasi*, penghapusan kata penghubung, dan *stemming*.

- *Cleaning* dilakukan untuk membersihkan teks dari elemen yang tidak relevan atau dapat mengganggu analisis, seperti *retweet, URL,* dan emotikon. Langkah ini bertujuan untuk mempertahankan hanya informasi teks yang penting.
- *Case Folding*, yaitu mengubah huruf kapital menjadi huruf kecil *(lowercase)*. Hal ini dilakukan agar tidak ada perbedaan dalam pemrosesan teks berdasarkan kapitalisasi huruf.

- *Tokenisasi* adalah proses mengurai kalimat menjadi kata-kata yang terpisah. Ini memungkinkan pemrosesan lebih lanjut pada tingkat kata dan membantu dalam analisis teks.
- Penghapusan kata penghubung *(stopword removal)* dilakukan untuk menghilangkan kata-kata yang umum tetapi tidak memberikan kontribusi signifikan pada pemahaman konten teks. Contohnya adalah kata penghubung seperti "dan", "atau", "di", dan sebagainya.
- S*temming* yang bertujuan untuk merubah kata-kata menjadi bentuk dasarnya atau kata dasar. Misalnya, kata "berlari" menjadi "lari", "bermain" menjadi "main", dan seterusnya. Hal ini dilakukan untuk mengurangi variasi kata yang serupa agar dapat menggambarkan makna yang sama. Dengan melakukan *preprocessing teks* ini, data menjadi lebih terstruktur dan siap untuk diolah lebih lanjut dalam tahap analisis.

2.2.5 *Term - Frequency Inverse Document Frequency (TF – IDF)*

Metode *Term Frequency Inverse Document Frequency (TF-IDF)* digunakan secara umum untuk menghubungkan kata-kata (*term*) dengan dokumen atau kalimat dengan memberikan bobot atau nilai pada setiap kata tersebut. Konsep utama dalam metode *TF-IDF* adalah menggabungkan frekuensi kata dalam sebuah dokumen *(Term Frequency)* dengan kebalikan frekuensi kata tersebut di seluruh dokumen *(Inverse Document Frequency)*. Untuk menghitung *TF* ada tiga cara yaitu [23] :

1. Menggunakan nilai biner pada kata-kat yang ada pada dokumen dan nilai 0 pada kata-kata yang tidak ada pada dokumen.
2. Menggunakan nilai frekuensi secara langsung pada kemunculan kata-kata untuk menjadi *TF*.
3. Menggunakan nilai pecahan term yang sudah dilakukan normalisasi, dengan rumus (2.1).

$$TF\ (t,d) = f_{t,d} \qquad (2.1)$$

Dimana :

(t,d) = dengan t adalah frekuensi kata yang muncul pada dokumen d.

Inverse Document Frequency (IDF) juga merupakan salah satu dari ekstarsi fitur pada dokumen yang dipre-processing, jika *TF* mencari banyaknya kata pada dokumen maka *IDF* mecari bobot dari setiap dokumen yang ada dengan mencari kata yang sama pada setiap dokumen. Dengan persamaannya menggunakan rumus (2.2) sebagai berikut :

$$IDF(t,d) = log\ \frac{N}{Df(t,d)} \qquad (2.2)$$

Dimana :

N = jumlah banyaknya dokumen

$Df(t,D)$ = jumlah banyaknya dokumen dalam dokumen d yang terdapat term t.

Term Frequency-Inverse Document Frequency (TF-IDF) adalah kombinasi dari *TF* dan *IDF* untuk menentukan indeks suatu dokumen dengan mengkombinasikan kedua situasi dari *TF* dan *IDF* hanya dengan menkalikan hasil dari *TF* dan *IDF* setiap *term*. Dengan persamaan rumus (2.3) sebagai berikut :

$$TF - IDF(t,d,D) = tf(t,d)\ x\ idf(t,D) \qquad (2.3)$$

2.2.6 *Naive Bayes*

Naive Bayes adalah metode sederhana dalam proses klasifikasi probabilitas yang mengacu pada Teori Bayes. Teori Bayes menyatakan bahwa probabilitas terjadinya suatu peristiwa dapat dihitung dengan mengalikan probabilitas intrinsik (berdasarkan data yang ada saat ini) dengan probabilitas bahwa peristiwa serupa akan terjadi di masa depan (berdasarkan pengetahuan yang diperoleh dari masa lalu). *Naive Bayes* merupakan algoritma pembelajaran probabilitas yang berasal dari teori Keputusan Bayesian. Algoritma ini mengasumsikan bahwa setiap fitur (kata atau atribut) dalam data adalah independen secara statistik, meskipun dalam kenyataannya ada keterkaitan antara fitur-fitur tersebut. *Naive Bayes* menggunakan rumus (2.4) sebagai berikut. |

$$P(H|\mathrm{X}) = \frac{P(H|\mathrm{X}).P(H)}{P(X)} \qquad (2.4)$$

Dimana :

X = Data dengan class yang belum diketahui.

H = Hipotesis data merupakan suatu class spesifik.

P(H|X) = Probabilitas hipotesis H berdasarkan 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 [24].

2.2.7 *Confusion Matrix*

Confusion matrix adalah sebuah tabel yang memberikan informasi tentang perbandingan antara hasil klasifikasi yang dilakukan oleh sistem (prediksi) dengan hasil klasifikasi yang sebenarnya. Tabel pada *confusion matrix* menunjukkan jumlah data uji yang diklasifikasikan dengan benar (*true positives* dan *true negatives*) dan jumlah data uji yang salah diklasifikasikan (*false positives dan false negatives*). *Confusion matrix* membantu dalam mengevaluasi performa model klasifikasi dengan memberikan gambaran tentang sejauh mana model tersebut mampu mengklasifikasikan data dengan benar. Dengan memeriksa elemen-elemen dalam *confusion matrix,* seperti akurasi, *presisi*, *recall,* dan *F1-score*, kita dapat memahami tingkat keberhasilan dan kegagalan model dalam melakukan klasifikasi. Tabel ini berisi empat kategori utama yaitu *True Positives* (TP), *True Negatives* (TN), *False Positives* (FP), dan *False Negatives* (FN). Berikut adalah penjelasan untuk masing-masing metrik evaluasi yang umum digunakan berdasarkan Confusion Matrix:

- Akurasi (*Accuracy*)

  Akurasi adalah ukuran sejauh mana model mampu melakukan klasifikasi dengan benar, yaitu berapa banyak data yang diklasifikasikan secara tepat (baik positif maupun negatif) dibandingkan dengan total data uji. Akurasi menggambarkan persentase keseluruhan prediksi yang benar, formula Akurasi (2.5) sebagai berikut.

  $$\text{Akurasi} = (TP + TN) / (TP + TN + FP + FN) \quad (2.5)$$

- Presisi (*Precision*)

Presisi mengukur sejauh mana hasil positif yang dihasilkan oleh model adalah relevan. Dalam konteks klasifikasi, presisi mengukur berapa banyak di antara hasil prediksi positif yang benar (TP) dibandingkan dengan total hasil prediksi positif (TP + FP), formula Presisi (2.6) sebagai berikut.

$$\text{Presisi} = \text{TP} / (\text{TP} + \text{FP}) \quad (2.6)$$

- *Recall* (Sensitivitas atau *True Positive Rate*)

  Recall mengukur sejauh mana model mampu menemukan atau mengidentifikasi keseluruhan data positif yang ada. Dalam konteks klasifikasi, recall mengukur berapa banyak di antara data positif yang berhasil diidentifikasi dengan benar (TP) dibandingkan dengan total data positif sebenarnya (TP + FN), formula *Recall* (2.7) sebagai berikut.

$$Recall = \text{TP} / (\text{TP} + \text{FN}) \quad (2.7)$$

- F1-Score

  F1-Score adalah harmonik rata-rata antara presisi dan *recall*. Metrik ini berguna ketika kita ingin mencari keseimbangan antara presisi dan *recall*. F1-Score menggabungkan informasi dari kedua metrik ini menjadi satu angka yang menyajikan kualitas keseluruhan dari model, formula F1-Score (2.8) sebagai berikut.

$$\text{F1} - \text{score} = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (2.8)$$

Saat menghadapi kasus dengan jumlah data positif dan negatif yang tidak seimbang, F1-Score seringkali menjadi metrik yang lebih baik daripada akurasi untuk mengevaluasi performa model. Hal ini karena akurasi dapat memberikan kesan model yang baik jika data negatif sangat dominan, tetapi sebenarnya model tersebut buruk dalam mengidentifikasi data positif yang penting. F1-Score dapat memberikan pandangan yang lebih realistis tentang kualitas klasifikasi model.