BAB II
TINJAUAN PUSTAKA

A. Penelitian Terdahulu

Sangat penting bagi penulis untuk memahami referensi dan hubungan antara penelitian sebelumnya dan saat ini, hal ini akan membantu mencegah duplikasi penelitian. Di samping itu, kajian literatur juga bermanfaat untuk memahami keberhasilan studi yang berpotensi berkontribusi pada kemajuan ilmu pengetahuan. Beberapa penelitian yang telah dilakukan dan dimanfaatkan sebagai dasar penelitian ini berkaitan dengan data dan metode yang digunakan. Rangkuman dari penelitian terdahulu ditunjukan oleh Tabel 2.1 sebagai berikut:

Tabel 2. 1 Tabel Penelitian Terdahulu

NO	Judul	Penulis, Tahun	Metode yang digunakan	Tujuan
1	Analisis Sentimen Pada Media Sosial Twitter Terhadap Tokoh Gus Dur Menggunakan Metode *Naïve Bayes* Dan *Support Vector Machine* (SVM)	(Salsabila, 2022)	Pada penelitian ini menggunakan *Support Vector Machine* (SVM) dan pendekatan *Naïve Bayes*	Memahami analisis sentimen Gus Dur di Twitter menggunakan *Support Vector Machine* (SVM) dan pendekatan *Naïve Bayes*

Tabel 2.1 Tabel Penelitian Terdahulu (lanjutan)

NO	Judul	Penulis, Tahun	Metode yang digunakan	Tujuan
2	Analisis Sentimen Masyarakat Terhadap Paylater Menggunakan Metode *Naive Bayes Classifier*	(Safira & Hasan, 2023)	*Naïve Bayes*	Memahami Analisis Sentimen Publik Metode Pengklasifikasi *Naïve Bayes* terhadap Paylater
3	Penerapan Algoritma *Naive Bayes* Untuk Analisis Sentimen Review Data Twitter BMKG Nasional	(Darwis et al., 2021)	Algoritma *Naïve Bayes*	Mengetahui Analisis Sentimen Review Data Twitter BMKG Nasional
4	Implementasi Algoritma *Naive Bayes* Terhadap Analisis Sentimen Opini Film Pada Twitter	(Fajar, 2018)	Algoritma *Naïve Bayes*	Mengetahui Analisis Sentimen Opini Film Pada Twitter

Tabel 2.1 Tabel Penelitian Terdahulu (lanjutan)

NO	Judul	Penulis, Tahun	Metode yang digunakan	Tujuan
5	Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (Dpr) Pada Twitter Menggunakan Metode *Naive Bayes* Classifier	(Putri et al., 2022)	Algoritma *Naïve Bayes*	Menganalisis Sentimen Kinerja DPR Pada Twitter Menggunakan *Naïve Bayes*
6	*Sentiment Analysis of Twitter Data*	(Wang et al., 2022)		Untuk memberikan Gambaran singkat dan hampir komprehensif teknik TSA dan bidang terkait.
7	*Machine learning and deep learning sentiment analysis models: case study on the sent-covid corpus of tweets in mexican spanish*	(Gomez-adorno et al., 2024)	*Naïve Bayes*	Untuk Analisis Sentimen Pembelajaran Mesin dan Pembelajaran Mendalam: Studi Kasus pada Kumpulan Tweet SENT-COVID dalam bahasa Spanyol Meksiko

Penelitian yang dilakukan oleh (Salsabila, 2022) Algoritma *Naïve Bayes* dan *Support Vector Machine* (SVM) digunakan dalam studi kuantitatif ini

tentang analisis sentimen untuk mengklasifikasikan opini tentang tokoh Gus Dur di Twitter. Alur kerja penelitian menjalani tahapan-tahapan SEMMA, yaitu Sample, Explore, Modify, Model, dan Asses. Pengumpulan sampel dilakukan dengan mengambil set data dari *tweet* yang di-crawl, diikuti dengan pemilihan atribut yang penting dalam tahap *explore*. Data kemudian diolah menjadi data terstruktur melalui proses *preproccesing* teks. Model untuk pelabelan diuji menggunakan masing-masing metode. Setelah dilabeli, data dikelompokkan dengan menggunakan kedua metode tersebut. Evaluasi hasil dilakukan dengan menerapkan *confusion matrix* dan *k-fold cross validation*. Akurasi yang didapat dari model Naïve Bayes ialah 78,36%, sedangkan *Support Vector Machine* menghasilkan akurasi 84,27% sehingga *Support Vector Machine* lebih unggul 5,91% dari Naïve Bayes. Di jejaring sosial X, Gus Dur dilihat dengan baik karena *Support Vector Machine* memeriksa 86 pikiran negatif kurang dari sentimen positif. Subjek karakter Gus Dur menghasilkan hasil akurasi yang baik dari *Naïve Bayes* dan *Support Vector Machine* (SVM), memungkinkan teknik ini digunakan untuk klasifikasi analisis sentimen pada data baru.

Penelitian yang dilakukan oleh (Safira & Hasan, 2023) Melakukan analisis sentimen menggunakan teknik Pengklasifikasi *TextBlob* dan *Naive Bayes* dari modul *TextBlob* bahasa pemrograman Python. Ada 405 poin data yang dikumpulkan melalui Twitter. 70,62% sentimen negatif, menurut hasil analisis sentimen menggunakan *Naive Bayes* Classifier, yang setara dengan 286 data. Selain itu, terdapat 22,72% sentimen positif, yang sama dengan 92 data, dan 6,67% sentimen netral yang diwakili oleh 27 data. Metode TextBlob juga menunjukkan lebih banyak sentimen negatif, mencapai 55,8% atau 226 data, diikuti oleh sentimen positif sebesar 33,09% atau 134 data, dan sentimen netral sebesar 11,11% atau 45 data. Bisa disimpulkan bahwa masyarakat tidak terlalu menyukai penggunaan paylater. Menggunakan *confusion matix* untuk menguji model menunjukkan bahwa algoritma klasifikasi *Naive Bayes* memiliki tingkat akurasi sebesar 91%, jauh lebih tinggi daripada *TextBlob* yang hanya mencapai 61%.

Berikutnya penelitian dari (Darwis et al., 2021) Pilihan pengklasifikasi Naïve Bayes adalah teknik klasifikasi data yang digunakan dalam penelitian ini. Untuk mengenali pernyataan yang mengandung pikiran positif, netral, atau negatif, metode ini dibangun menggunakan data internal dari internet dan X. Penentuan itu dibuat melalui proses klasifikasi. Penelitian telah dimasukkan ke dalam *fined grained sentiment analysis,* yang mana memeriksa kalimat komentar secara mendalam. Data tersebut diolah menggunakan teknik text mining, Setelah itu, dibagi menjadi tiga kelas: netral, negatif, dan positif. Algoritma *Naive Bayes* digunakan untuk tujuan klasifikasi. Pemisahan tersebut membantu untuk memudahkan pengguna mengidentifikasi sudut pandang netral, negatif, dan positif. Hasil pengujian klasifikasi menggunakan pendekatan *Naive Bayes* memiliki akurasi sebesar 69,97%.

Penelitian yang dilakukan oleh (Fajar, 2018) Dengan begitu banyaknya opini yang tercantum di Media sosial X, penting untuk mengkategorikan mereka berdasarkan sentimen yang tersebar, sehingga dapat dengan mudah melihat arah sentimen yang berlaku dari film tersebut, baik yang menguntungkan atau tidak menguntungkan. Dalam penelitian ini, algoritma *Naive Bayes* diterapkan. Hasil eksperimen menunjukkan bahwa sistem dapat melakukan analisis sentimen dengan tingkat *accuracy* mencapai 90%, *precission* mencapai 92%, dengan nilai *recall* dan *f-measure* juga mencapai 90%.

B. Landasan Teori

1. Analisis Sentimen

Analisis sentimen adalah teknik untuk memahami dan memproses data teks secara otomatis untuk mengidentifikasi sentimen yang ada dalam opini (Sari & Wibowo, 2019). Opini yang dimaksud mengarah pada opini terhadap sebuah objek yang dapat berupa topik, produk, layanan, organisasi, individu, acara, maupun masalah. Salah satu kelebihan dari analisis sentimen adalah kapasitasnya untuk mengurangi waktu dan tenaga saat bekerja dengan banyak data dalam penelitian. Analisis sentimen memiliki tujuan untuk mendapatkan

informasi tentang kadar positif, negatif, atau netral dari suatu opini. Analisis tersebut kemudian dijadikan pertimbangan penting dalam pengambilan keputusan. Ada beberapa contoh yang merupakan penerapan analisis sentimen seperti dalam bisnis, penting untuk memahami reputasi produk baru di kalangan masyarakat guna dapat meningkatkan citra produk tersebut. Dalam bidang politik, perlu mengetahui popularitas seorang tokoh untuk mendapatkan informasi yang lebih baik tentang tokoh tersebut dan contoh program seperti vaksinasi, penanganan hukum, pemilihan presiden, pemilihan kepala daerah dan lainnya digunakan untuk meningkatkan pemerintahan.

Secara umum, analisis sentimen dikerucutkan ke dalam lima fase yaitu pengumpulan data, *preprocessing*, *featur selection*, *classification*, dan *evaluation*. Data acak dapat diubah menjadi data yang dapat dipahami menggunakan analisis sentimen. Analisis sentimen memberikan manfaat yang besar dengan memberikan evaluasi dan pemikiran yang berharga di berbagai bidang, analisis sentimen bisa dipakai untuk meneliti kejadian, ucapan, dan komentar yang dapat diperdebatkan. Temuan analisis sentimen juga dapat membantu pemerintah, bisnis, dan tokoh publik memilih tindakan mereka selanjutnya (Natasuwarna, 2020).

2. X

X merupakan platform jejaring sosial dan *mikroblogging* online yang memungkinkan pengguna untuk mengirim dan membaca pesan teks hingga 280 karakter. Pada tanggal 7 November 2017, panjang karakter diperpanjang menjadi 280 karakter, yang secara umum disebut sebagai "*tweet* " X didirikan pada bulan Maret 2006 oleh Jack Dorsey dan diluncurkan pada bulan Juli 2006. Sejak awal dirilis, X telah tumbuh menjadi salah satu dari 10 situs yang paling sering dikunjungi di dunia maya dan disebut sebagai "pesan singkat di internet ". X saat ini merupakan salah satu platform media sosial yang paling diminati oleh pengguna online. Berbeda dengan media sosial lain, X membatasi penggunanya yang hanya bisa melakukan penulisan hingga 280 karakter. Hal ini memungkinkan pengguna untuk menyampaikan informasi atau sudut pandang secara lebih ringkas dengan membuat teks atau *tweet* yang

dilihat pengguna X lebih pendek, lebih padat, dan lebih jelas (Paputungan & Jacobus, 2021).

3. Naïve Bayes

Teknik penambangan data yang disebut Naïve Bayes dapat memperkirakan probabilitas berdasarkan pengalaman sebelumnya dan digunakan untuk mengklasifikasikan hal-hal menggunakan statistik dan probabilitas. *Support Vector Machine* (SVM), *K-Nearest Neighbor* (KNN), *Artificial Neural Network* (ANN*), Trees Gradient Boosted* (TGB), dan *Random Trees* (RT) adalah contoh klasifikasi pembelajaran yang diawasi, sedangkan Decision Tree, Logistic Regression, dan Kernel Regression adalah contoh regresi. (Roihan, Sunarya, dan Rafika 2020). Salah satu metode yang menggunakan teknik probabilistik adalah *Naive Bayes*, di mana dua fitur dalam kumpulan data yang sama tidak terkait (Insan et al., 2023). Adapun persamaan dari *teorema bayes* ditunjukan pada persamaan (1) sebagai berikut:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

Keterangan :

X : Data yang belum diketahui.

H : Hipotesis data X merupakan suatu class spesifik.

P(H|X) : Probabilitas hipotesis H berdasarkan kondisi X (*Posterior Probability*).

P(H) : Probabilitas hipotesis H (*Prior Probability*).

P(X|H) : Probabilitas X berdasarkan Hipotesis H.

P(X) : Probabilitas X.

4. Metode Data Mining

Data mining adalah proses memperoleh informasi yang berguna dari database besar untuk diekstraksi untuk membuat informasi baru yang mendukung pengambilan keputusan. Data mining terdiri dari beberapa langkah yang melibatkan interaksi langsung pengguna atau menggunakan basis pengetahuan. Berikut adalah tahap-tahap dalam tahap mining:

a. Pembersihan Data

Yaitu tahap menghilangkan data yang mengganggu, tidak valid, atau tidak relevan merupakan bagian dari proses pembersihan data.

b. Integrasi Data

Penggabungan data dari berbagai sumber ke dalam basis data baru adalah yang dimaksud dengan integrasi data. Proses ini membolehkan pengguna mengakses dan menganalisis data dari segala sumber dengan lebih lancar dan efisien.

c. Seleksi Data

Database hanya dicari untuk informasi yang dapat dianalisis karena tidak semua data digunakan sehingga diperlukan seleksi data

d. Transformasi Data

Data tersebut diubah atau digabungkan menjadi format yang dapat digunakan untuk proses mining. Prosedur ini disebut *pre-processing* data, yang merupakan langkah penting dalam menjamin keakuratan dan kualitas tinggi data yang digunakan dalam analisis data mining.

e. Proses Mining

Yaitu proses untuk melihat pola dan struktur tersembunyi dalam data.

f. Evaluasi Pola (Pola *Pattern*)

Tujuannya adalah untuk mengungkap pola menarik dalam basis informasi yang ditemukan. Basis informasi ini kemudian dapat digunakan untuk menjawab pertanyaan kompleks tentang data dan membuat keputusan yang lebih baik.

5. *Preprocesing*

Proses memilih data mana yang akan diproses untuk setiap dokumen dikenal sebagai *preprocessing*. Untuk menghapus kata-kata yang berlebihan dari data, tahap pra-pemrosesan itu penting. (Nurian, 2023) tahap preprocesing adalah:

a. *Case folding*

Selama langkah *preprocessing* yang dikenal sebagai "*Case folding*", setiap huruf dalam dokumen diubah menjadi huruf kecil.

b. *tokenizing*

Tokenizing merupakan proses preprocessing yang melakukan pemecahan kata pada kalimat. Umumnya dalam kalimat, kata-kata dipisahkan menggunakan spasi untuk membantu proses tokenisasi.

c. *Filtering*

Langkah ini dilakukan sebagai upaya untuk menghapus data dengan kesalahan ketik, kesalahan, atau ketidaklengkapan.

d. *Stemming*

Steaming yaitu Proses menemukan kata dasar dilakukan dengan menghapus semua imbuhan yang melekat pada kata.

e. *Normalization*

Mengubah kata-kata dalam data set ke bentuk baku yang lebih formal menurut KBBI.

6. *Confusion Matrix*

Salah satu alat untuk menentukan nilai akurasi hasil analisis sentimen adalah matriks kebingungan. Data aktual dan yang diantisipasi membentuk *confusion matrix*. *Confusion matrix* memungkinkan kita untuk menghitung *precision* dan *recall* untuk setiap kelas (faradhillah 2016).

Nilai lain yang dapat digunakan untuk mengevaluasi metode yang dikembangkan dapat diperoleh dengan menggunakan keempat variabel ini, yaitu (Normawati & Prayogi, 2021) :

a. *Accuracy*. menjelaskan metrik untuk mengukur tingkat akurasi model. Nilai akurasi dapat diperoleh menggunakan persamaan (2) sebagai berikut:

$$Accuracy = \frac{TP + TN}{Jumlah\ data} \quad (2)$$

b. *Precision*. Kemampuan model untuk mengidentifikasi kelas target dibandingkan dengan semua hasil yang diprediksi kelas disebut sebagai presisi. Contohnya, Kapasitas untuk memperkirakan Semua estimasi kelas positif memiliki data nyata positif. Penilaian yang akurat dapat diperoleh dengan persamaan (3) sebagai berikut:

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

c. *Recall*. Kapasitas model untuk memanggil kelas target menggunakan semua datanya. Misalnya, berdasarkan data aktual positif keseluruhan kumpulan data, model memperkirakan data aktual positif. Nilai penarikan dapat diperoleh dengan persamaan (4) sebagai berikut:

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

d. *F1 – score. F1 – score* merupakan rata-rata *precission* dan *recall*, nilai *F1 - skor* dapat diperoleh dengan persamaan (5) sebagai berikut:

$$F1 - score = 2 * \frac{Precision * Recall}{Precicion + Recall} \quad (5)$$

1) *True Positive* (TP) adalah jumlah data positif nyata yang diidentifikasi oleh model sebagai hasil dari prediksi positif.
2) *False Positive* (FP) adalah jumlah data negatif nyata yang diidentifikasi oleh model sebagai hasil dari prediksi positif.
3) *False Negative* (FN) adalah jumlah data positif nyata yang diidentifikasi oleh model sebagai hasil dari prediksi negatif.
4) *True Negative* (TN) adalah jumlah data negatif nyata yang diidentifikasi oleh model sebagai hasil dari prediksi negatif.