BAB II

TINJAUAN PUSTAKA

2.1 Teori – Teori Dasar

Sub bab ini membahas teori-teori dasar yang berfungsi sebagai landasan konseptual dalam penelitian ini. Teori-teori tersebut digunakan untuk memberikan pemahaman yang mendalam terhadap permasalahan yang dikaji serta menjadi dasar dalam melakukan analisis terhadap data yang diperoleh. Dengan mengacu pada teori-teori yang relevan, penelitian ini memperoleh pijakan ilmiah yang jelas dan sistematis dalam menjelaskan fenomena yang diteliti.

2.2.1. NLP (*Natural Language Processing*)

Natural Language Processing (NLP) atau pemrosesan bahasa alami merupakan salah satu cabang Artificial Intelligence (AI) yang mempelajari pembuatan sistem untuk menerima masukan bahasa alami manusia. Dalam perkembangannya, NLP berusaha untuk mengubah bahasa alami komputer (bit dan byte) menjadi bahasa alami manusia yang dapat kita mengerti. NLP merupakan ilmu dasar yang dapat dijadikan jembatan untuk membuat komunikasi antara mesin dengan manusia dengan memproses bahasa, baik lisan atau tulisan yang digunakan oleh manusia dalam komunikasi sehari-hari (Soyusiawaty, 2023). Komponen *Natural Language Processing* (NLP):

1. *Dictionary*/Kamus/Leksikon

   Kamus adalah sebuah rujukan yang menerangkan makna kata-kata. Kamus mendaftarkan kata berdasarkan abjad dan informasi tentang bagaimana mereka

dipakai. Setiap entry dalam kamus merepresentasikan informasi tentang kata benda hidup, kata benda mati, kata sifat, kata kerja, kata keterangan, kata depan, kata sandang, kata petunjuk dan kata ganti yang diketahui untuk sistem pengolahan bahasa alami.

2. Parser

   Parser merupakan elemen yang paling menentukan dalam *Natural Language Processing* (NLP). Parser adalah sepenggal software yang dapat menganalisis input kalimat secara sintaktik. Parser melakukan idenifikasi tiap-tiap kata dan kemudian membuat peta kata-kata tersebut dalam struktur yang disebut pohon parser. Pohon parser menunjukkan makna dalam semua kata dan bagaimana cara menggabungkan kata-kata tersebut. Parser juga dapat mengidentifikasi prasa kata kerja, prasa kata benda dan selanjutnya memilah-milah kedalam elemen-lemen yang lain.

3. *Knowledge Representation System*/Sistem Representasi Pengetahuan

   Tahap ini digunakan untuk menganalisis output parser guna menentukan maknanya. Didalam sistem berisi fakta-fakta, teori, pemikiran dan hubungan antara satu dengan yang lainnya.

4. *Output Translator*

   Suatu terjemahan yang merepresentasikan sistem pengetahuan dan melakukan langkah-langkah yang bisa berupa jawaban atas bahasa alami atau output khusus yang sesuai dengan program komputer lainnya.

Kesulitan dalam *Natural Language Processing* (NLP)

a. *Ambiguity*

Ambiguity artinya mempunyai makna lebih dari satu. Keambiguan ini dapat menimbulkan keraguan atau ketidakjelasan dalam kalimat yang diucapkan atau ditulis. Contohnya kata “bisa” yang bisa jadi memiliki arti “racun” atau “dapat”. Hal lain yang dapat menjadi ambigu yaitu simbol huruf dan tanda baca (simbol titik tidak selalu berfungsi sebagai akhir kaliat, tetapi dapat menjadi bagian dari singkatan). Contohnya “Ir., Dr., S.T, dan masih banyak lagi”.

b. Jumlah kosa kata (*vocabulary*) dalam bahasa alami sangat besar.

c. Text segmentation Sebagai contoh beberapa bahasa tulisan seperti Chinese, Japanese, dan Thai tidak memiliki baasan antara satu kata dengan yang lain.

d. *Word sense disambiguation*

   Yaitu kata-kata yang mempunyai lebih dari satu arti.

e. *Syntactic ambigulity*

   Yaitu terdapat banyaknya kemungkinan pohon parsing untuk satu kalimat.

2.2.2. Text Mining

Text mining merupakan suatu teknik atau metode analisis yang digunakan untuk menggali dan mengekstraksi informasi berharga dari jumlah besar teks atau dokumen. Dalam teknik ini, algoritma dan pendekatan statistik digunakan untuk memproses, mengorganisasi, dan menyajikan data teks dalam bentuk yang dapat dimengerti dan dianalisis oleh manusia atau sistem komputer. Selain itu, *text mining* juga dikenal sebagai data mining teks atau penemuan pengetahuan dari database tekstual. Berdasarkan definisi dari buku "*The Text Mining Handbook*", *text mining* bisa diartikan sebagai suatu proses dimana seorang pengguna berinteraksi dengan

sekumpulan dokumen menggunakan alat analisis yang merupakan komponen komponen dari data mining (Firdaus & Firdaus, 2021).

Analisis sentimen adalah suatu proses dalam *text mining* yang menggunakan algoritma data mining untuk mengklasifikasikan data tidak terstruktur dan menghasilkan informasi mengenai sentimen secara efisien. Tujuan dari *text mining* adalah untuk mendapatkan informasi yang berharga dari sekumpulan dokumen. Oleh karena itu, sumber data yang digunakan dalam *text mining* berupa teks yang tidak terstruktur atau setidaknya semi terstruktur. Beberapa tugas khusus yang dilakukan dalam text mining antara lain adalah pengkategorisasian dan pengelompokkan teks (Nurhuda et al., 2016).

Dalam memberikan solusi, *text mining* mengadopsi dan mengembangkan berbagai teknik dari bidang lain, seperti Data Mining, Information Retrieval, Statistik dan Matematik, Machine Learning, Linguistik, *Natural Language Processing*, dan Visualisasi. Kegiatan riset dalam text mining meliputi ekstraksi dan penyimpanan teks, *preprocessing* konten teks, pengumpulan data statistik, serta indexing dan analisis sentiment (Manullang et al., 2023).

2.2.3. Analisis Sentimen

Analisis sentimen merupakan salah satu cara untuk mengetahui pendapat orang banyak terhadap perilaku sesuatu seperti layanan publik, isu, kinerja pemerintahan atau hal lainnya. Analisis sentimen ini dilakukan untuk melihat kecenderungan opini masyarakat terhadap isu-isu yang berkembang terutama di media sosial. Analisis sentimen ini merupakan suatu pengklasifikasian dengan mengekstraksi pendapat, emosi, dan evaluasi seseorang yang tertulis dalam sebuah

pembicaraan mengenai topik tertentu dengan memanfaatkan *Natural Language Processing* (Aziz, 2022). Menurut Zulfa & Winarko, Analisis sentimen adalah sebuah riset komputasional dari opini, sentimen dan emosi yang diekspresikan secara tekstual (Lestari & Anggraeni, 2021). Oleh karena itu, analisis sentimen merupakan salah satu cabang dalam penelitian *text mining* yang melakukan klasifikasi dalam sebuah dokumen teks (Gazali Mahmud et al., 2023).

Tujuan dari analisis sentimen adalah untuk memahami dan mengukur reaksi emosional atau pendapat orang terhadap suatu topik, produk, layanan, atau peristiwa tertentu. Data yang telah terkumpul akan diolah melalui analisis agar dapat memperoleh pandangan yang mendalam sehingga kesimpulan dapat diambil. Hasil analisis sentimen dapat berupa sentimen positif, negatif, atau netral.

Menurut Pang dan Lee (2008), analisis sentimen memadukan teknik linguistik, statistik, dan pembelajaran mesin untuk mendeteksi kecenderungan emosional dalam suatu teks. Sentimen dalam konteks ini dapat diekspresikan secara eksplisit maupun implisit, dan sering kali memiliki nuansa yang mempersulit proses klasifikasi, seperti ironi, sarkasme, dan ambiguitas makna.

Analisis Sentimen dapat dibagi menjadi 3 level, yaitu :

a) Level Dokumen: Pada level ini, analisis sentimen bertujuan untuk mengklasifikasikan keseluruhan dokumen ke dalam kelas positif, netral, atau negative.
b) Level Kalimat: Pada level ini, analisis sentimen bertujuan untuk menentukan sentimen positif atau negatif dari suatu kalimat dengan mempertimbangkan susunan kata dalam kalimat tersebut.

c) Level Aspek: Pada level ini, analisis sentimen bertujuan untuk menentukan sentimen positif, negatif, atau netral berdasarkan atribut dari suatu entitas. Dalam penelitian ini, level aspek diterapkan untuk mengkategorikan ulasan sesuai dengan aspek yang telah ditentukan (Sormin et al., 2017)

Dalam konteks analisis sentimen pada dialog film, pendekatan ini menjadi lebih kompleks karena teks tidak selalu eksplisit menunjukkan emosi. Struktur bahasa dalam dialog juga cenderung informal, kontekstual, dan sering kali melibatkan ekspresi budaya atau humor yang tidak langsung.

Tantangan utama dalam analisis sentimen antara lain: Ketidakjelasan makna (*ambiguity*), Sarkasme atau ironi, Ketergantungan konteks, Ketidakseimbangan data (kelas sentimen yang tidak seimbang). Meskipun begitu, dengan kemajuan dalam teknik NLP dan algoritma klasifikasi seperti Naive Bayes, SVM, atau Deep Learning, tingkat akurasi dalam mendeteksi sentimen semakin meningkat, terutama dalam domain yang spesifik seperti dialog film.

2.2.4. Dialog Film

Dialog film merupakan komponen penting dalam naskah film yang merepresentasikan komunikasi verbal antara karakter. Dialog tidak hanya menyampaikan informasi, tetapi juga membangun karakter, mengungkap emosi, dan menciptakan dinamika naratif dalam alur cerita. Menurut McKee (1997), dialog dalam film memiliki fungsi utama sebagai alat ekspresi karakter, penggerak cerita, dan penyampai tema yang bersifat implisit.

Berbeda dengan bentuk teks lain seperti berita atau ulasan, dialog film cenderung lebih natural dan kontekstual, mencerminkan percakapan sehari-hari

yang penuh nuansa emosi, jeda, pengulangan, dan penggunaan bahasa informal. Oleh karena itu, menganalisis sentimen dalam dialog film memerlukan pendekatan khusus karena emosi tidak selalu diungkapkan secara eksplisit. Misalnya, kalimat "Ya udah, terserah kamu…" bisa bermakna netral, tapi dalam konteks bisa membawa nuansa kesal atau kecewa.

Dalam konteks pengolahan bahasa alami (NLP), dialog film termasuk dalam jenis unstructured text yang harus diolah melalui tahapan *preprocessing* sebelum dianalisis. Misalnya, dialog harus dibersihkan dari tanda baca, dijadikan bentuk dasar kata, dan dikonversi menjadi vektor fitur sebelum dapat diproses oleh algoritma klasifikasi seperti Complement Naive Bayes. Studi oleh (Walker et al., n.d.) menunjukkan bahwa dialog film juga bisa dianalisis untuk mengukur ekspresi emosional dan karakterisasi tokoh. Penelitian tersebut menggunakan korpus subtitle film sebagai data untuk mengidentifikasi pola sentimen antar karakter.

2.2.5. Teori Himpunan

Teori himpunan merupakan fondasi penting dalam matematika dan memiliki keterkaitan erat dengan konsep klasifikasi dalam algoritma Naive Bayes. Himpunan didefinisikan sebagai kumpulan objek atau elemen yang memiliki sifat tertentu dan dapat ditentukan secara eksplisit. Dalam konteks Naive Bayes, elemen-elemen fitur dan kelas dalam proses klasifikasi dapat dipandang sebagai anggota dari himpunan tertentu (Darwanto et al., 2020).

Operasi-operasi dasar dalam himpunan seperti union (∪), intersection (∩), dan difference (−), digunakan untuk memodelkan relasi antar fitur dan kelas. Sebagai contoh, irisan antara himpunan fitur tertentu dengan himpunan kelas

memberikan gambaran probabilitas bersyarat yang menjadi inti dari perhitungan Naive Bayes. Prinsip inklusif-eksklusif dan hukum aljabar himpunan juga menjadi alat bantu dalam memahami probabilitas gabungan dan eksklusif antar kejadian (Siregar et al., 2024).

Penerapan teori himpunan dalam studi ekonomi dan sosial membuktikan bahwa konsep ini bersifat universal dan dapat diadaptasi dalam berbagai bentuk analisis, termasuk untuk pengambilan keputusan berbasis data sebagaimana dilakukan oleh algoritma Naive Bayes (Zulfa et al., 2024).

2.2.6. Teori Probabilitas

Teori Probabilitas merupakan cabang dari statistika yang mempelajari kemungkinan terjadinya suatu peristiwa dalam ruang sampel tertentu. Dalam konteks analisis data dan pengambilan keputusan berbasis data, teori probabilitas memainkan peranan penting dalam memodelkan ketidakpastian serta mengukur keyakinan terhadap suatu kejadian. Probabilitas didefinisikan sebagai rasio antara banyaknya kejadian yang diharapkan dengan jumlah seluruh kemungkinan kejadian dalam ruang sampel. Secara matematis, jika suatu kejadian A merupakan bagian dari ruang sampel S, maka probabilitas $P\ (A)$ terjadi sesuai persamaan P1.

$$P(A) = \frac{n(A)}{n(S)} \qquad \text{(P1)}$$

di mana $n\ (A)$ adalah jumlah elemen dalam kejadian A, dan $n\ (S)$ adalah jumlah total elemen dalam ruang sampel S.

Salah satu teori fundamental dalam probabilitas adalah Teorema Bayes (*Bayes' Theorem*). Teorema ini memberikan dasar matematis untuk menghitung

probabilitas bersyarat, yaitu probabilitas suatu kejadian A terjadi dengan syarat bahwa kejadian B telah terjadi.

Teorema Bayes digunakan secara luas dalam berbagai bidang, terutama dalam metode klasifikasi statistik seperti algoritma Naive Bayes, di mana klasifikasi dilakukan berdasarkan perhitungan probabilitas bersyarat dari fitur-fitur yang diamati terhadap kelas tertentu (Soyusiawaty, 2023).

2.2.7. Naïve Bayes

Naïve Bayes Classifier (NBC) adalah metode klasifikasi yang berdasarkan probabilitas dan Teorema Bayesan dengan asumsi bahwa setiap variable X bersifat bebas (independence). Dengan kata lain NBC mengasumsikan bahwa keberadaan sebuah atribut (variable) tidak ada kaitannya dengan keberadaan atribut (variable) yang lain (Kurniawan et al., n.d.). Keuntungan menggunakan metode naïve bayes adalah metode ini hanya membutuhkan jumlah data latih yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses klasifikasi. Metod naïve bayes bisa bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan.

Perhitungan perbandingan antara term pada data uji dengan setiap kelas yang ada dengan menggunakan persamaan P2.

$$P(A|B) = \frac{P(B|A)xP(A)}{P(B)} \qquad \text{(P2)}$$

dengan:

A : Hipotesis data merupakan class spesifik.

B : Data dengan kelas yang masih belum diketahui.

P(A|B) : Probabilitas hipotesis berdasarkan kondisi.

P(A) : Probabilitas hipotesis.

P(B|A) : Probabilitas berdasarkan kondisi pada hipotesis.

P(B) : Probabilitas B.

2.2.8. Complement Naïve Bayes

Complement Naïve Bayes (CNB) merupakan pengembangan lainnya dari metode Naïve Bayes Classifier dengan memodifikasi persamaan akhir yang digunakan pada metode MNB. Pada model CNB, alih alih menghitung probabilitas dari suatu kelas menggunakan probabilitas kemunculan fitur berdasarkan kelas tersebut, CNB menghitung probabilitas dari suatu kelas dengan menggunakan probabilitas kemunculan fitur pada kelas lainnya. Complement Naïve Bayes (CNB) dikembangkan sebagai perbaikan dari Multinomial Naïve Bayes (MNB), khususnya untuk mengatasi masalah ketidakseimbangan data atau performa yang buruk pada klasifikasi teks. CNB menggunakan informasi dari komplemen kelas (dokumen yang bukan dari kelas target) untuk menghitung bobot fitur, sehingga cenderung lebih akurat. Berdasarkan modifikasi yang dilakukan algoritma CNB dipercaya lebih baik dalam melakukan proses klasifikasi berdasarkan dataset dalam keadaan tidak seimbang (*imbalance dataset*) (Budiman et al., 2021).

Persamaan Complement Naïve bayes dapat dilihat pada (P3).

$$score(c,d) = \sum_{i=1}^{n} xi.log\left(\frac{P(w_i|\bar{c})}{\Sigma_j P(w_j|\bar{c})}\right) - log\big(P(c)\big) \qquad \text{(P3)}$$

dengan:

c: Kelas target yang sedang dievaluasi (misalnya: Konflik, Persahabatan, *Multiverse*).

d: Dokumen uji, yaitu teks atau kalimat yang akan diklasifikasikan.

w_i: Kata ke-i yang terdapat dalam dokumen d.

x_i: Jumlah kemunculan kata w_i dalam dokumen d.

$P(w_i|\bar{c}\,)$: Probabilitas kata w_i berdasarkan kelas selain c (komplemen dari kelas target).

$\Sigma_j P(w_j|\bar{c}\,)$: Jumlah total probabilitas dari semua kata dalam kelas komplemen $\bar{c}$, digunakan untuk normalisasi.

$P\,(\,c\,)$: Prior probabilitas kelas c, yaitu proporsi jumlah data dalam kelas c terhadap seluruh data.

2.2 Tinjauan Penelitian Terdahulu

Bagian ini menyajikan kajian terhadap penelitian terdahulu yang relevan untuk memperkuat dasar teoritis dan argumen penelitian. Tinjauan dilakukan dengan membandingkan pendekatan, metode, serta temuan dari studi sebelumnya guna menemukan celah penelitian (research gap). Referensi penelitian terdahulu dirangkum dalam Tabel 2.1 agar memudahkan pembaca memahami perkembangan penelitian sejenis serta menegaskan posisi dan orisinalitas penelitian ini. Melalui kajian ini, peneliti dapat memastikan bahwa penelitian yang dilakukan memiliki kontribusi baru. Selain itu, tinjauan ini juga berfungsi sebagai pijakan untuk mengarahkan strategi penelitian agar lebih tepat sasaran.

Tabel 2. 1 Penelitian Terdahulu

No	Penulis	Judul	Metode	Hasil Penelitian
1.	Nadhif Sanggara Fathullah, dkk (2020)	Analisis Sentimen Terhadap Rating dan Ulasan Film dengan menggunakan Metode Klasifikasi Naïve Bayes dengan Fitur Lexicon-Based	klasifikasi Naïve Bayes dengan fitur Lexicon-Based	Hasil penelitian menunjukkan bahwa hasil pengujian, nilai akurasi, precision, dan recall dengan pemilihan fitur berupa penghapusan stopword mendapatkan hasil masing-masing 0,9, 0,9, dan 0,9, sedangkan nilai akurasi, precision, dan recall dengan pemilihan fitur berupa Lexicon-Based memiliki hasil masing-masing 1, 1, dan 1.
2.	Yuni Nurtikasari, dkk (2022)	Analisis Sentimen Opini Masyarakat Terhadap Film Pada Platform Twitter Menggunakan Algoritma Naive Bayes	Naïve Bayes	Hasil penelitian menunjukkan bahwa pengujian dengan confusion matrix dengan tools orange didapatkan hasil rata – rata nilai akurasi 0.65% dan nilai presisi sebesar 0.67%, dan recall sebesar 0.65%, dan persentase netral 0.83% pada klasifikasinya.
3.	Okta Ihza Gifari, dkk (2022)	Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine	TF IDF dan SVM	Hasil penelitian menunjukkan bahwa algoritma TF IDF dan SVM dapat digunakan untuk kasus review film dengan nilai Accuracy 85%, nilai Precision 100%, nilai Recall 70%, dan nilai F1-Score sebesar 82%.
4.	Muhammad Al-Fajr Ramadhani, dkk (2024)	Klasifikasi Sentimen Opini terhadap Film Kartini Menggunakan Naive Bayes pada Platform X	Naive Bayes Classifier	Hasil penelitian menunjukkan bahwa Hasil akurasi terbaik dari dataset emosi yang didapatkan adalah 98% dan akurasi terbaik dari dataset sentimen yang didapatkan adalah 86%.
5.	Adinda Cahya Kamilla, dkk (2024)	Analisis Sentimen Film Agak Laen Dengan Kecerdasan Buatan: Text Mining Metode Naïve Bayes Classifier	Naïve Bayes Classifier	Hasil penelitian menunjukkan bahwa setelah dilatih dan dievaluasi menggunakan KFold cross validator dengan k=5, model menunjukkan akurasi sebesar 78%. Evaluasi menunjukkan model memiliki precision sebesar 85.51%, recall sebesar 85.57%, dan f1-score sebesar 82.71%. Dengan demikian, metode klasifikasi Naïve Bayes Classifier memiliki potensi yang baik dalam menganalisis sentimen ulasan film di Twitter.

Sumber: Diolah dari berbagai sumber (Cahya Kamilla et al., 2024; Fathullah et al., 2020; Gifari et al., 2022; Nurtikasari et al., 2022; Ramadhani et al., 2024).

Dalam beberapa tahun terakhir, penelitian yang terkait analisis sentimen pada konten film baik dari segi ulasan penonton maupun tanggapan di media social semakin banyak dilakukan, khususnya dengan penerapan algoritma pembelajaran mesin seperti Naïve Bayes. Penelitian dari (Fathullah et al., 2020), mengevaluasi ulasan dan rating film menggunakan metode Naïve Bayes dengan tambahan fitur Lexicon-Based. Hasil penelitian menunjukkan nilai akurasi, precision, dan recall yang mencapai 100% saat menggunakan fitur tersebut. Meski begitu, objek kajiannya masih sebatas ulasan pengguna, bukan percakapan atau dialog dalam film itu sendiri.

Berbeda dengan (Nurtikasari et al., 2022), penelitian ini mengkaji opini masyarakat di Twitter terkait film, tetap menggunakan algoritma Naïve Bayes. Tetapi, hasilnya kurang optimal, dengan akurasi rata-rata hanya sekitar 65%. Ini menunjukkan data dari media sosial, yang biasanya tidak terstruktur dan penuh dengan kata-kata tidak baku, cukup sulit dianalisis secara akurat.

Sedangkan penelitian dari (Gifari et al., 2022) mencoba pendekatan lain dengan menggabungkan metode TF-IDF dan algoritma Support Vector Machine (SVM). Mereka berhasil mendapatkan akurasi yang lebih baik, yakni 85%, bahkan precision-nya mencapai 100%. Namun, fokus penelitian ini masih berada pada ulasan teks film, bukan pada dialog film.

Selanjutnya, Penelitian oleh (Ramadhani et al., 2024), penelitian ini lebih fokus pada opini publik terhadap film "Kartini" menggunakan Naïve Bayes. Penelitian ini mencatat akurasi yang sangat tinggi pada data emosi, mencapai 98%.

Meski menjanjikan, penelitian ini terbatas pada satu film saja dan belum menyentuh analisis percakapan antar karakter dalam film.

Sementara itu, (Cahya Kamilla et al., 2024) menganalisis sentimen terhadap film "Agak Laen" yang sedang ramai di Twitter, menggunakan metode Naïve Bayes. Dengan validasi silang, mereka mencapai akurasi 78% dan f1-score sebesar 82,71%. Meski hasilnya cukup baik, penelitian ini masih meneliti data media sosial, bukan pada teks dialog film yang memiliki struktur lebih formal dan naratif.

Dari beberapa penelitian tersebut, Penelitian ini menghadirkan pendekatan berbeda dengan menganalisis dialog film sebagai sumber data utama, yang cenderung lebih ekspresif, kontekstual, dan terstruktur dibandingkan ulasan penonton atau opini di media sosial. Algoritma Complement Naïve Bayes dipilih karena kemampuannya menangani dataset tidak seimbang, yang kerap menjadi kendala dalam klasifikasi sentimen. Pendekatan ini diharapkan menjadi alternatif yang lebih sesuai untuk data dialog film serta memberikan kontribusi dalam pengembangan analisis sentimen yang selama ini lebih berfokus pada opini pengguna.