BAB II

TINJAUAN PUSTAKA

2.1 Definisi & Teori

2.1.1 *Natural Language Processing*

Natural Language Processing (NLP) atau yang biasa disebut pemrosesan Bahasa alami merupakan salah satu bidang dari *Artificial Intelligence* (AI) yang berfokus pada pengembangan sistem yang mampu memahami dan mengolah input berupa Bahasa alami yang digunakan manusia (Soyusiawaty, 2023). Dengan demikian, memungkinkan terciptanya jalur komunikasi yang lebih mudah dan efisien antara manusia dan komputer (Hasibuan & Heriyanto, 2022) dengan memproses Bahasa baik baik itu lisan maupun tulisan yang digunakan oleh manusia dalam sehari-hari. Dalam NLP, juga penting untuk memperhatikan pemahaman pada struktur dan karakteristik Bahasa itu sendiri yang mencakup pemilihan kosa kata, cara penyususnan kata menjadi kalimat yang bermakna, makna setiap kalimat dan lain sebagainya (Soyusiawaty, 2023)

Dengan kemajuan teknologi yang semakin meningkat, NLP menjadi salah satu bidang yang paling inovatif dalam ranah AI. NLP tidak hanya digunakan untuk memproses dan menganalisis teks, tetapi juga memungkinkan komputer untuk memahami, membentuk serta menanggapi Bahasa manusia dengan tingkat kecanggihan yang semakin tinggi (Tarumingkeng, 2024). NLP telah digunakan secara luas di berbagai bidang, seperti analisis sentimen, analisis media sosial,

chatbot, penerjemahan bahasa, ekstraksi informasi, dan berbagai aplikasi lainnya (Hasibuan & Heriyanto, 2022).

2.1.2 *Text Mining*

Text mining termasuk kedalam cabang dari data mining (Andrasthea & Februariyanti, 2024) yang berfungsi untuk melakukan ekstraksi data dalam bentuk teks dalam jumlah besar secara berkala. Data yang digunakan pada proses ini umumnya berupa koleksi dokumen yang tidak terstruktur, yang perlu untuk dikelompokkan untuk mempermudah pengolahan secara efisien (Azzahra, 2023). Selain itu, *text mining* juga berfungsi untuk mengidentifikasi kata-kata yang menggambarkan isi dokumen, yang kemudian digunakan dalam proses analisis keterkaitan antar dokumen (Andrasthea & Februariyanti, 2024). Pada proses ini, *text mining* melibatkan beberapa proses diantaranya sebagai berikut:

a. Data *Cleaning*
b. *Tokenizing*
c. *Stopword*
d. *Stemming*

2.1.3 Analisis Sentimen

Analisis sentimen termasuk dalam cabang *Natural Language Processing* (NLP) atau *Text Mining* yang merupakan bagian dari Ranah *Machine Learning* (ML) (Purnamasari et al., 2023). Analisis sentimen sendiri merupakan salah satu metode yang digunakan untuk melakukan analisis opini, emosi (Sari et al., 2020) , mengekstrak data dan mengolah data secara otomatis, sehingga juga biasa disebut dengan *opinion mining* (Fauziah, 2023). Analisis sentimen telah menunjukkan

pengaruh yang besar hampir di seluruh bidang, salah satunya di media sosial. Semakin berkembangnya teknologi, semakin besar pula masyarakat menggunakan media sosial dan semakin meningkat pula opini masyarakat sehingga analisis sentimen menjadi sangat penting untuk perusahaan untuk mengetahui respon dari masyarakat terhadap produk atau jasa mereka. Hal ini dikarenakan analisis sentimen dapat mengkategorikan opini masyarakat menjadi positif, netral dan negatif (Purnamasari et al., 2023) sehingga dapat membantu untuk meningkatkan pelayanan serta kualitas (Azzahra, 2023) dan kekuatan *branding* (Purnamasari et al., 2023).

Dari banyaknya manfaat analisis sentimen diberbagai bidang, terdapat pula tantangan yang perlu ditangani seperti terkait big data dan bahasa. Menurut Gouthami & Hedge (2021) tantangan analisis sentimen yang terkait pada big data yaitu mencakup a. pengumpulan data, b. pemrosesan awal data, c. penyimpanan dan analisis data, d. kecepatan data besar dan e. keragaman data. Sedangkan tantangan yang terkait Bahasa meliputi a. kurangnya korpus dan kamus yang tersedia, b. gaya penulisan yang berbeda, c. arti/makna kata yang berbeda, dan d. berbagai Bahasa sangat kontekstual. Selain itu adanya kata yang ambigu, kalimat sarkasme, data *quality* dan emoji sehingga sulit untuk terdeteksi.

Menurut Sajid (2013) analisis sentimen bisa dibedakan menjadi beberapa tingkatan seperti berikut:

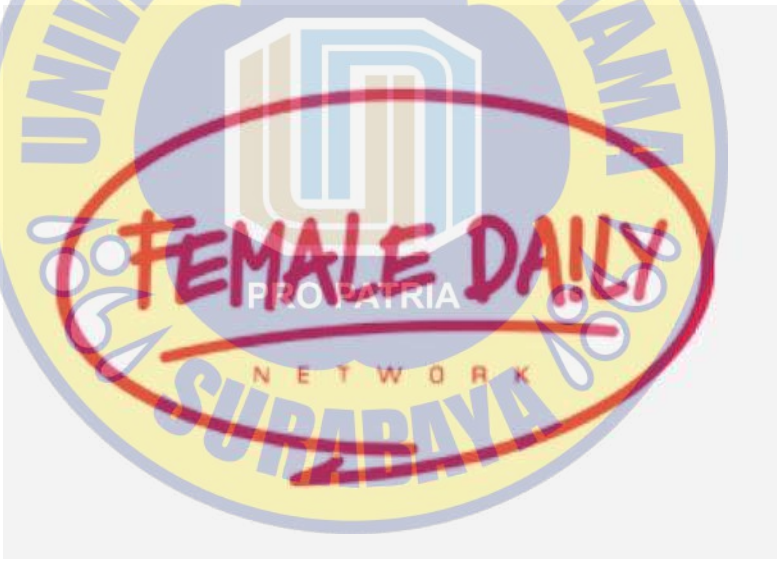
a. *Multimodal sentiment analysis*. Analisis jenis ini menggunakan indicator visual dan pendengaran sebagai aspek yang digunakan untuk mempresentasikan berbagai macam sentimen contohnya seperti ekspresi

wajah dan nada suara. Data yang bisa digunakan untuk analisis jenis ini yaitu video, audio dan teks.

b. *Aspect-based sentiment analysis*. Analisis jenis ini menggunakan metode NLP dalam proses analisis dan digunakan untuk menganalisis emosi, opini dari produk maupun layanan.

c. *Multilingual sentiment analysis*. Analisis jenis ini menggunakan teks dengan berbagai Bahasa untuk menganalisis karena mempunyai tata Bahasa, sintaksis dan kosa kata yang berbeda (Purnamasari et al., 2023).

2.1.4 Female Daily

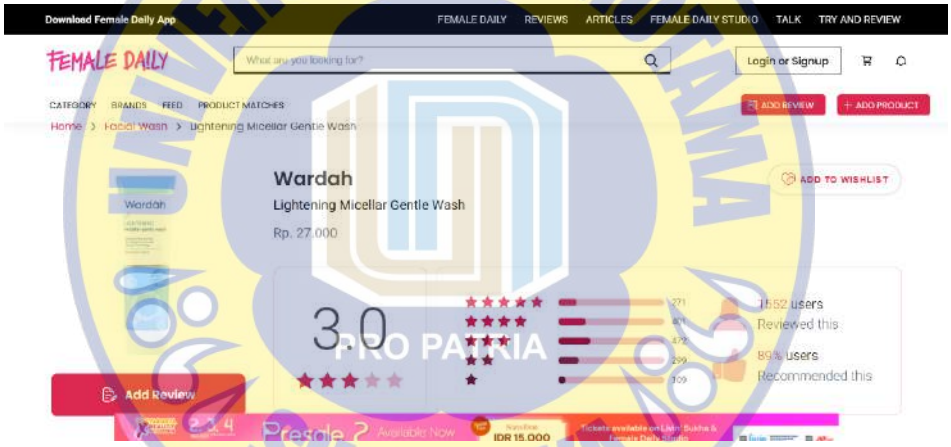


Gambar 2. 1 Logo Female Daily

Sumber: Kompasiana.com

Gambar 2.1 merupakan logo Female Daily yang merupakan platform digital sekaligus komunitas yang berfokus pada perempuan dan kecantikan (Nabila, 2022). Female Daily berdiri sejak tahun 2005 yang didirikan oleh Hanifa Ambadar dan Affi Assegaf yang bermula dari blog pribadi (Kamal, 2021) sehingga berkembang

menjadi komunitas online yang memfasilitasi pengguna untuk berbagi pengalaman pribadi (Azzahra, 2023) dengan konten fashion dan kecantikan (Kamal, 2021). Menurut Wardani (2017) menjelaskan bahwa Female Daily berada di bawah pengelolaan Pt Daily Dinamika Kreasi dengan jumlah viewers mencapai 7,8 juta perbulan dan memiliki 18 ribu thread forum sejak dilaporkan (Nabila, 2022). Pengguna dapat dengan mudah mengakses layanan Female Daily melalui aplikasi berbasis smartphone, IOS serta situs web resmi seperti pada Gambar 2.2 (Kamal, 2021).



Gambar 2. 2 Website Female Daily

Sumber: femaledaily.com

Female Daily sendiri memiliki tujuan untuk memberdayakan perempuan Indonesia dengan menyediakan berbagai informasi yang dibutuhkan, maka dari itu Female Daily menawarkan berbagai macam fitur yang meliputi ulasan produk, workshop seputar kecantikan, tutorial, pendapat para ahli, ulasan konsumen serta forum diskusi online (Kamal, 2021). Membaca review dari konsumen kini menjadi langkah penting sebelum membeli produk sehingga dengan banyaknya fitur yang

ditawarkan dan ulasan dari konsumen yang terpercaya mampu membantu mempermudah konsumen dalam memilih produk yang sesuai dengan kebutuhan (Nabila, 2022).

2.1.5 Pembersih Wajah/ *Facial Wash*

Pembersih wajah atau biasa disebut *facial wash* merupakan salah satu jenis perawatan kulit dasar yang selalu ada dalam rutinitas perawatan kulit yang berbahan dasar yang memiliki kandungan *surfaktan* (Nabila, 2022). *Facial wash* merupakan produk dengan formulasi yang lembut dan ringan yang berperan dalam menjaga kebersihan kulit. Produk ini menjadi salah satu pilihan praktis dan ekonomis dalam mengatas masalah jerawat (Nirmala et al., 2021). Hal ini karena *facial was*h dapat menghilangkan kotoran, minyak, debu serta sisa make up. Selain itu pembersih wajah juga berfungsi untuk mencegah berbagai macam masalah kulit seperti jerawat, komedo dan lainnya. Namun setiap produk pembersih wajah memiliki manfaat yang berbeda sesuai dengan kandungan di dalamnya (Larasati et al., 2024).

2.1.6 Himpunan

Teori himpunan merupakan salah satu cabang ilmu matematika dalam bidang analisis yang berkembang pada era modern yang ditemukan oleh Georg Ferdinand Ludwig Phillip Cantor pada akhir abad ke-19 (Junarti, 2023). Istilah himpunan dalam matematika berasal dari kata “set” yang berasal dari Bahasa Inggris. Himpunan juga bisa disebut dengan kata kelas, gugus dan kelompok. Secara singkat himpunan dapat diartikan sebagai sekumpulan objek (Darwanto et al., 2020). Himpunan digunakan untuk membangun hampir segala aspek dari matematika serta merupakan sumber dari mana semua matematika diturunkan (Junarti, 2023).

Himpunan merupakan sekumpulan objek yang sifatnya dapat diartikan sebagai koleksi benda-benda tertentu yang dianggap sebagai satu kesatuan (Junarti, 2023). Objek yang ada pada himpunan antara lain disebut dengan elemen, unsur, atau anggota (Darwanto et al., 2020). Sehingga studi tentang struktur pada himpunan elemen-elemennya sangatlah penting (Junarti, 2023).

2.1.7 Probabilitas

Probabilitas merupakan cabang dari matematika yang mempelajari tentang ukuran kemungkinan terjadinya suatu peristiwa dalam kehidupan sehari-hari. Kata "probabilitas" sendiri berasal dari Bahasa Inggris "probably" yang berarti kemungkinan. Dengan demikian, probabilitas dapat diartikan sebagai kemungkinan suatu kejadian yang dapat terjadi dalam kondisi yang tidak pasti. Dalam konteks ini, probabilitas juga dikenal sebagai teori peluang. Penggunaan probabilitas sangat penting dalam pengambilan keputusan, terutama karena kehidupan ini penuh dengan ketidakpastian. Dengan mengetahui tingkat kemungkinan terjadinya suatu peristiwa, seseorang dapat membuat keputusan yang lebih baik dan tepat sasaran (Chamdani, 2022).

Terdapat tiga pendekatan utama yang digunakan untuk mendefinisikan probabilitas dan menentukan nilainya, yaitu:

a. Pendekatan Klasik**:** Berdasarkan pada jumlah seluruh kemungkinan yang dapat terjadi dari suatu kejadian dengan asumsi bahwa setiap kemungkinan memiliki peluang yang sama.
b. Pendekatan Frekuensi Relatif (Objektif)**:** Menentukan probabilitas berdasarkan proporsi dari frekuensi terjadinya suatu kejadian dalam percobaan atau

observasi berulang.

c. Pendekatan Subjektif**:** Probabilitas ditentukan berdasarkan keyakinan atau penilaian pribadi yang dinyatakan dalam bentuk derajat kepercayaan (Pane & Silvanita, 2022)

Selain pendekatan tersebut, probabilitas memiliki tiga elemen penting, yaitu:

1) Percobaan (*Experiment*)**:** Proses atau aktivitas yang dapat menghasilkan suatu hasil tertentu.
2) Hasil (*Outcome*)**:** Merupakan satu kemungkinan hasil dari suatu percobaan.
3) Peristiwa (*Event*)**:** Merupakan kumpulan dari satu atau lebih hasil yang terjadi dalam suatu percobaan.

Dalam probabilitas, dikenal pula konsep probabilitas bersyarat, yang menyatakan peluang terjadinya suatu peristiwa dengan syarat bahwa peristiwa lain telah terjadi sebelumnya. Notasi untuk probabilitas bersyarat adalah P(A|B), yang berarti peluang A terjadi dengan syarat B telah terjadi sesuai persamaan P1.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{(P1)}$$

Salah satu bagian penting dari teori probabilitas adalah *Teorema Bayes*, yang digunakan untuk menghitung peluang dari suatu hipotesis berdasarkan bukti atau data yang diamati. *Teorema Bayes* merupakan dasar dari statistika Bayes dan memiliki banyak penerapan, termasuk dalam bidang ekonomi mikro, ilmu pengetahuan, teori permainan, hukum, dan dunia medis (Pane & Silvanita, 2022).

Teorema Bayes juga banyak digunakan dalam sistem pakar atau sistem berbasis kecerdasan buatan. Dalam sistem ini, nilai probabilitas dari hipotesis dan

bukti (*evidence*) digunakan untuk menghasilkan keputusan berdasarkan informasi yang diperoleh dari objek yang didiagnosis (Chamdani, 2022).

2.1.8 *Naïve Bayes*

Naïve bayes merupakan metode yang klasifikasi yang berbasis pada pendekatan probabilitas yang dikenal sebagai *Teorema Bayes* yang digunakan untuk memprediksi kejadian masa depan berdasarkan pada data sebelumnya yang dikembangkan oleh Thomas Bayes ilmuwan dari Inggris (Hasanah et al., 2022). *Naïve Bayes* juga merupakan salah satu algoritma yang banyak dimanfaatkan dalam klasifikasi terutama data dengan bentuk teks serta telah menunjukkan hasil yang cukup baik dalam penerapannya. *Naïve Bayes* dikenal sebagai algoritma yang memiliki kemampuan klasifikasi yang cepat, selain itu juga terbukti efektif dan efisien bahkan saat diterapkan pada analisis skala besar (Setyaningsih et al., 2023). Sehingga membuat *Naïve Bayes* menjadi salah satu algoritma klasifikasi yang paling popular pada bidang *machine learning* dan data mining. *Teorema Naïve Bayes* terjadi sesuai dengan persamaan P2.

$$P(C|X) = \frac{P(X|C).P(C)}{P(X)} \qquad \text{(P2)}$$

Keterangan:

P(C|X) : Probabilitas kelas C, diberikan fitur X

P(C) : Probabilitas prior kelas C

P(X|C) : Probabilitas fitur X muncul dalam kelas C

P(X) : Probabilitas fitur X

2.1.9 *Gaussian Naïve Bayes*

Gaussian Naïve Bayes merupakan salah satu dari jenis algortima *Naïve Bayes* (Cahyaningrum et al., 2022) yang digunakan apabila fiturnya bertipe kontinu (Pratama et al., 2022). *Gaussian Naïve Bayes* merupakan metode yang mengasumsikan distribusi data menggunakan pola normal *(gaussian)*, dengan menghitung nilai rata-rata dan standar *deviasi* dari data pelatihan. Pada metode ini, probabilitas dari setiap nilai input dihitung untuk masing-masing kelas. Jika data yang digunakan bersifat kontinu, maka perhitungan rata-rata dan standar *deviasi* dilakukan terhadap masing-masing nilai input (Nugroho et al., 2023). *Gaussian Naïve Bayes* terjadi sesuai dengan persamaan P3.

$$P(X = x|C = c) = \frac{1}{\sqrt{2\pi\sigma}} e \frac{-(x-\mu)2}{2\sigma 2} \quad \text{(P3)}$$

Keterangan:

X= variabel

C= kelas

μ= mean

σ= deviasi stander

2.2 Penelitian Terdahulu

Pada penelitian ini, peneliti merujuk pada sejumlah penelitian terdahulu yang telah membahas topik-topik serupa, baik dari segi metode analisis, pendekatan klasifikasi sentimen, maupun objek kajian yang sejenis, guna memperkuat landasan teoritis serta memberikan pemahaman yang lebih komprehensif terhadap konteks permasalahan yang diangkat dalam studi ini. Oleh karena itu, daftar penelitian

terdahulu yang dijadikan acuan telah disusun secara sistematis dan disajikan secara rinci dalam Tabel 2.1 sebagai bentuk pemetaan literatur yang jelas.

Dari Tabel 2.1 terlihat bahwa sebagian besar penelitian terdahulu berfokus pada perbandingan antar model klasifikasi sentimen, dan dari empat tabel tersebut menunjukkan bahwa metode SVM memiliki akurasi yang lebih tinggi dibandingkan metode *Naïve Bayes*, meskipun selisih nilai akurasinya tidak terlalu signifikan. Namun, pada tabel ketiga justru ditemukan hasil berbeda, dimana metode *Naïve Bayes* unggul dalam hal akurasi, yang mengindikasikan bahwa performa model sangat dipengaruhi oleh karakteristik dataset, representasi fitur, serta teknik praproses yang digunakan. Temuan ini memperlihatkan bahwa tidak ada satu metode yang secara mutlak lebih baik untuk semua kasus, sehingga pemilihan algoritma harus mempertimbangkan konteks dan jenis data yang dianalisis.

Berdasarkan pertimbangan tersebut, peneliti mengusung penelitian ini dengan fokus pada analisis sentimen produk pembersih wajah menggunakan metode Gaussian Naïve Bayes, yang dikenal memiliki proses prediksi cepat, efisien dalam penggunaan sumber daya, sederhana dalam implementasi, serta mampu memberikan performa yang baik pada dataset besar dengan fitur numerik seperti kata kunci tertentu. Diharapkan, penelitian ini tidak hanya memberikan kontribusi dalam membandingkan efektivitas metode yang digunakan, tetapi juga memberikan wawasan baru mengenai penerapan *Gaussian Naïve Bayes* pada domain produk kecantikan. Selain itu, hasil penelitian ini diharapkan dapat menjadi rujukan bagi pengembang sistem analisis sentimen untuk menentukan metode yang tepat sesuai karakteristik data yang mereka miliki.

Tabel 2. 1 Penelitian Terdahulu

No	Penulis	Judul	Data Penelitian	Metode	Hasil Penelitian
1.	Cindy Nada Adela, dkk (2024)	Analisis Ulasan Pengguna Aplikasi Seabank Dengan Support Vector Machine Dan Naïve Bayes	Ulasan Aplikasi Seabank dari Google Play Store	SVM dan Naïve Bayes	Hasil penelitian menunjukkan bahwa model SVM lebih efektif dalam mengklasifikasikan sentimen ulasan pengguna aplikasi Seabank daripada model Naïve Bayes.
2.	Difa Durrotun Nada (2022)	Perbandingan Analisis Sentimen Mengenai BPJS Pada Media Sosial Twitter Menggunakan Naïve Bayes Classifier Dan Support Vector Machine (SVM)	Ulasan Twitter	SVM dan Naïve Bayes Classifier	Hasil penelitian menunjukkan bahwa SVM kernel RBF dengan parameter C=1000 γ=100 memilki performa ketepatan klasifikasi yang paling baik dibandingkan Naïve Bayes Classifier dan SVM Linear dengan hasil rata-rata 97,1%, 92,5% dan 86,7%.
3.	Hermanto, dkk (2024)	Perbandingan Algoritma Klasifikasi Analisis Sentimen Pengguna Aplikasi Getcontact Dalam Pencegahan Penipuan Online.	Ulasan aplikasi getcontact di Google Play	SVM dan Naïve Bayes	Hasil penelitian menunjukkan bahwa Naïve Bayes lebih unggul dalam mengkalsifikasikan komentar pengguna aplikasi Getcontact di google play sebagai komentar positif dan negatif.
4.	M. Rafli Saputra, dkk (2025)	Analisis Sentimen Twitter Terhadap Konflik Di Papua Menggunakan Perbandingan Naïve Bayes Dan SVM	Twitter	SVM dan Naïve Bayes	Hasil penelitian menunjukkan bahwa algoritman Naïve Bayes memiliki akurasi sebesar 95% sedangkan SVM sebesar 99% .
5.	Sugeng Setyabudi, dkk (2024)	Analisis Sentimen Penilaian Pengguna Marketplace Lazada Dengan Metode Naïve Bayes Dan Support Vector Machine	Google Play Store	SVM dan Naïve Bayes	Hasil penelitian menunjukkan bahwa dari hasil pengujian SVM memiliki performa lebih baik dengan akurasi 75%, precision 74%, recall 86%, dan F1-score 79%, dibandingkan dengan Naïve Bayes memiliki akurasi sebesar 72%, precision 75%, recall 76%, dan F1-score 76% pada sentimen positif.

Sumber: Hasil Penelitian Diolah Kembali