

BAB II

TINJAUAN PUSTAKA

2.1 Kajian Pustaka

Analisis sentimen telah diterapkan dalam berbagai konteks, mulai dari menilai ulasan aplikasi hingga memantau respons di media sosial. Pada penelitian – penelitian sebelumnya, banyak yang menggunakan metode Naive Bayes. Berikut penelitian terdahulu yang menurut penulis mempunyai hubungan dengan penelitian yang dilakukan.

Pertama, penelitian [7] berjudul “App Review Analisis Sentimen Aplikasi Shopee di Google Play Store menggunakan Algoritma Naïve Bayes” oleh Dany Pratmanto dan kolega pada tahun 2020 bertujuan membantu manajemen Shopee menganalisis ulasan. Hasilnya menunjukkan bahwa algoritma Naïve Bayes mencapai akurasi klasifikasi 96,667%, *precision* 100%, *recall* 93,33%, dan AUC 1,00, yang termasuk dalam klasifikasi *excellent*. Penelitian ini menganalisis 200 ulasan pengguna Shopee di *Google Play Store* [7]. Sementara itu, penelitian ini melakukan analisis sentimen pengguna aplikasi Riliv yang berfokus pada kesehatan mental, berbeda dengan studi sebelumnya yang berfokus pada Shopee.

Kedua, penelitian [8] “Analisis Sentimen Opini Publik mengenai Covid-19 pada Twitter menggunakan Metode Naive Bayes dan KNN” oleh Muhammad Syarifuddin pada tahun 2020 menilai pandangan masyarakat terhadap Covid-19, apakah positif atau negatif, dan membandingkan akurasi antara Naive Bayes dan KNN. Hasil menunjukkan Naive Bayes lebih akurat (63,21%) dibandingkan KNN (58,10%). Analisis sentimen di Twitter mengungkapkan opini positif lebih dominan dengan 610 opini positif dan 488 opini negatif. Naive Bayes juga memiliki *precision* lebih tinggi untuk opini positif (66,40%) dibandingkan opini negatif (58,94%) [8]. Penelitian ini berbeda karena meneliti sentimen terhadap aplikasi Riliv melalui ulasan pengguna.

Ketiga, penelitian “Analisis Sentimen Masyarakat terhadap Layanan Konsultasi Dokter menggunakan Algoritma Naive Bayes” oleh Neng Wardani dan Adhitia Erfina tahun 2021 mengevaluasi tiga aplikasi dokter online: Alodokter, Halodoc, dan Klik Dokter. Naive Bayes mencapai akurasi tertinggi pada Klik Dokter (98,57%) [9]. Penelitian sebelumnya berfokus pada layanan konsultasi dokter, berbeda dengan penelitian ini yang fokus pada aplikasi Riliv.

Keempat, penelitian “Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid-19 Menggunakan Algoritma Naïve Bayes Classifier” oleh Winda Yulita dan tim (2021) menemukan mayoritas masyarakat (60,3%) memiliki respon positif terhadap vaksin Covid-19 [10]. Penelitian ini bertujuan memahami pandangan masyarakat mengenai vaksin Covid-19, sementara penelitian ini fokus pada ulasan pengguna aplikasi Riliv.

Kelima, penelitian “Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter” oleh Dwi Normawati dan Surya Allit Prayogi (2021) menganalisis opini publik tentang berita kontroversial Ahok di Twitter menggunakan Naïve Bayes. Hasilnya menunjukkan akurasi 60%, presisi 100%, dan recall 50% [11]. Penelitian tersebut menganalisis sentimen berbasis teks di Twitter, sementara penelitian ini mengevaluasi respons pengguna terhadap aplikasi Riliv.

Keenam, penelitian “Analisis Sentimen Aplikasi E-Government Pada Google Play Menggunakan Algoritma Naive Bayes” oleh Artanti Inez Tanggraeni dan Melkior N. N. Sitokdana (2022) mengkategorikan ulasan aplikasi E-Government menjadi positif atau negatif dengan akurasi 89% [12]. Penelitian ini dan penelitian sebelumnya menggunakan metode Naive Bayes, tetapi penelitian ini fokus pada aplikasi Riliv.

Ketujuh, Penelitian “Perbandingan Akurasi Metode Lexicon Based dan Naive Bayes Classifier Pada Analisis Sentimen Pendapat Masyarakat Terhadap Aplikasi Investasi Pada Media Twitter” oleh Fitrah Amaliah dan I Kadek Dwi Nuryana (2022) menemukan Naive Bayes lebih akurat (78%) daripada metode *Lexicon Based* (67%) [13]. Penelitian sebelumnya meneliti perbandingan antara dua metode analisis sentimen, yaitu *Lexicon Based* dan *Naive Bayes Classifier*, sementara

penelitian ini fokus pada penggunaan metode Naive Bayes untuk mengevaluasi respon pengguna terhadap aplikasi Riliv.

Tabel 2.1 Penelitian Terdahulu

No	Judul	Penulis	Tahun	Studi Kasus	Metode / Algoritma	Hasil	Perbandingan dengan Penelitian yang Dilakukan
1	App Review Analisis Sentimen Aplikasi Shopee di <i>Google Play Store</i> menggunakan Algoritma Naïve Bayes	Dany Pratmanto, Rousyati Rousyati, Fanny Fatma Wati, Andrian Eko Widodo, Suleman Suleman dan Ragil Wijianto	2020	Data ulasan pengguna aplikasi Shopee.	Metode Naïve Bayes	Penelitian klasifikasi sentimen menggunakan metode Naïve Bayes menunjukkan hasil yang sangat memuaskan. Algoritma ini menghasilkan tingkat akurasi klasifikasi sebesar 96,667%, dengan <i>presisi</i> 100% dan <i>recall</i> 93,33%. Algoritma ini mampu mengidentifikasi dengan tepat 28 data <i>review</i> positif dan 30 data berita negatif. Hanya terdapat 2 data <i>review</i> negatif yang diklasifikasikan sebagai positif, dan 0 data berita positif diklasifikasikan sebagai negatif.	Berbeda dengan penelitian sebelumnya yang berfokus pada analisis sentimen ulasan pengguna Shopee, penelitian ini memfokuskan pada ulasan pengguna aplikasi Riliv yang berkaitan dengan kesehatan mental. Penelitian terdahulu memilih aplikasi Shopee sebagai studi kasus, sedangkan penelitian ini mengangkat aplikasi Riliv sebagai objek penelitiannya.

No	Judul	Penulis	Tahun	Studi Kasus	Metode / Algoritma	Hasil	Perbandingan dengan Penelitian yang Dilakukan
2	Analisis Sentimen Opini Publik mengenai Covid-19 pada Twitter menggunakan Metode Naive Bayes dan KNN	Muhammad Syarifuddin	2020	Opini publik mengenai COVID-19 yang diungkapkan melalui <i>platform</i> Twitter.	Metode Naïve Bayes dan KNN	Analisis sentimen yang dilakukan pada penelitian ini menunjukkan bahwa mayoritas opini publik terkait Covid – 19 tergolong positif. Dua metode klasifikasi yang digunakan, yaitu Naive Bayes dan KNN, menghasilkan akurasi yang berbeda. Naive Bayes menunjukkan performa yang lebih unggul dengan akurasi sebesar 63,21%, dibandingkan dengan KNN. Melalui penelitian ini, penulis berharap dapat berkontribusi dalam memperluas pemahaman masyarakat umum tentang Covid – 19.	Berbeda dengan penelitian sebelumnya yang memeriksa sentimen opini publik terkait Covid – 19 di Twitter, penelitian ini difokuskan pada analisis sentimen terhadap aplikasi Riliv berdasarkan ulasan pengguna yang ditemukan di <i>Google Play Store</i> dan <i>App Store</i> .

No	Judul	Penulis	Tahun	Studi Kasus	Metode / Algoritma	Hasil	Perbandingan dengan Penelitian yang Dilakukan
3	Analisis Sentimen Masyarakat terhadap Layanan Konsultasi Dokter menggunakan Algoritma Naive Bayes	Neng Resti Wardani dan Adhitia Erfina	2021	Ulasan Masyarakat terhadap Layanan Konsultasi Dokter	Naïve Bayes	Penelitian ini menunjukkan bahwa Klik Dokter memiliki akurasi tertinggi (98,57%), diikuti Halodoc (82,86%) dan Alodokter (62,86%) dalam klasifikasi aplikasi kesehatan menggunakan algoritma Naive Bayes. Penulis menduga rendahnya akurasi Alodokter disebabkan oleh penyimpanan data yang tidak andal, sehingga banyak pengguna memberikan komentar negatif. Peningkatan jumlah data latih dapat meningkatkan akurasi, namun hal ini juga dapat berakibat sebaliknya jika kata-kata dalam data ulasan bias atau bermakna ganda.	Berbeda dengan penelitian sebelumnya yang berfokus pada analisis sentimen untuk membantu pengembang meningkatkan aplikasi, penelitian ini bertujuan untuk mengidentifikasi perbandingan antara ulasan positif dan negatif yang diberikan pengguna aplikasi Riliv di <i>Google Play Store</i> dan <i>App Store</i> .

No	Judul	Penulis	Tahun	Studi Kasus	Metode / Algoritma	Hasil	Perbandingan dengan Penelitian yang Dilakukan
4	Analisis Sentimen terhadap Opini Masyarakat tentang Vaksin Covid-19 menggunakan Algoritma <i>Naïve Bayes Classifier</i>	Winda Yulita, Eko Dwi Nugroho, Muhammad Habib Algifari	2021	Opini masyarakat tentang vaksin Covid-19	Algoritma Naïve Bayes Classifier	Mayoritas masyarakat Indonesia (60,3%) menyambut baik kebijakan vaksinasi Covid – 19. Hanya sebagian kecil yang menunjukkan penolakan (5,4%) atau sikap netral (34,4%). Analisis sentimen menggunakan algoritma Naive Bayes mencapai tingkat akurasi yang tinggi, yaitu 93%.	Penelitian sebelumnya berfokus pada pemahaman pandangan masyarakat tentang vaksin Covid-19 dan menggolongkan sentimen ke dalam kategori positif, negatif, atau netral. Namun, penelitian ini berbeda karena menargetkan perbandingan antara ulasan positif, netral, dan negatif yang diberikan oleh pengguna terhadap aplikasi Riliv.

No	Judul	Penulis	Tahun	Studi Kasus	Metode / Algoritma	Hasil	Perbandingan dengan Penelitian yang Dilakukan
5.	Implementasi <i>Naive Bayes Classifier</i> dan <i>Confusion Matrix</i> pada Analisis Sentimen Berbasis Teks pada Twitter	Dwi Normawati dan Surya Allit Prayogi	2021	Opini masyarakat terhadap berita kontroversi Ahok di Twitter	Metode yang diterapkan dalam penelitian ini adalah <i>Naive Bayes Classifier</i> dan <i>Confusion Matrix</i> .	Penelitian ini menerapkan metode Naive Bayes untuk menganalisis sentimen di Twitter dan berhasil memberikan informasi penting tentang opini publik. Metode ini terbukti akurat (60%) dan mampu mengidentifikasi sentimen dengan baik (<i>recall</i> 50%). Dalam studi kasus dengan 5 data latih dan 3 data uji, Naive Bayes berhasil memprediksi tingkat analisis sentimen dengan tingkat keberhasilan yang tinggi.	Penelitian sebelumnya fokus pada analisis sentimen berbasis teks pada <i>platform</i> Twitter, sedangkan penelitian ini fokus pada mengidentifikasi perbandingan antara ulasan positif, netral, dan negatif terhadap aplikasi Riliv di <i>Google Play Store</i> dan <i>App Store</i> .

No	Judul	Penulis	Tahun	Studi Kasus	Metode / Algoritma	Hasil	Perbandingan dengan Penelitian yang Dilakukan
6	Analisis Sentimen Aplikasi E-Government pada <i>Google Play</i> Menggunakan Algoritma Naive Bayes	Artanti Inez Tanggraen, Melkior N. N, Sitokdana	2022	Analisis sentimen pengguna terhadap aplikasi E-Government yang tersedia di <i>Google Play Store</i>	Algoritma Naive Bayes	Hasil penelitian menunjukkan bahwa pengguna lebih banyak memberikan ulasan positif (407) dibandingkan negatif (237). Hal ini mendukung penelitian sebelumnya yang menunjukkan kepuasan pengguna terhadap aplikasi. Penggunaan metode Naive Bayes pada 577 data pelatihan dan 65 data uji menghasilkan klasifikasi yang baik dengan akurasi 89%, <i>presisi</i> 83%, dan <i>recall</i> 87%.	Penelitian sebelumnya bertujuan untuk menganalisis sentimen pengguna terkait pengalaman pengguna menggunakan aplikasi tersebut dan memberikan wawasan bagi pengembangan lebih lanjut. Sedangkan penelitian ini bertujuan untuk mengidentifikasi perbandingan antara ulasan positif dan negatif pengguna untuk memahami masalah dan kekurangan yang perlu diperbaiki. Selain naive bayes, penelitian sebelumnya juga menggunakan metode <i>K-Nearest Neighbors</i> (KNN).

No	Judul	Penulis	Tahun	Studi Kasus	Metode / Algoritma	Hasil	Perbandingan dengan Penelitian yang Dilakukan
7.	Perbandingan Akurasi Metode <i>Lexicon Based</i> dan <i>Naive Bayes Classifier</i> pada Analisis Sentimen Pendapat Masyarakat terhadap Aplikasi Investasi pada Media Twitter	Fitrah Amaliah dan I Kadek Dwi Nuryana	2022	Analisis sentimen pendapat masyarakat terhadap aplikasi investasi pada media Twitter	Metode <i>Lexicon Based</i> dan <i>Naive Bayes Classifier</i>	Penelitian ini mengembangkan metode klasifikasi berbasis kosa kata dan Naive Bayes untuk menganalisis sentimen 698 tweet. Setelah pra-pemrosesan, 147 tweet dianalisis. Metode berbasis kosa kata mencapai akurasi 67% dengan persentase polaritas positif tertinggi 64,63%. Metode Naive Bayes mencapai akurasi 78% dengan persentase polaritas positif tertinggi 53,74%. Perbandingan menunjukkan bahwa Naive Bayes lebih akurat (78%) daripada metode berbasis <i>Lexicon Based</i> (67%).	Metode yang digunakan dalam penelitian sebelumnya adalah perbandingan antara metode <i>Lexicon Based</i> dan <i>Naive Bayes Classifier</i> , sedangkan penelitian ini menggunakan metode Naive Bayes untuk mengidentifikasi perbandingan ulasan positif dan negatif. Hasil penelitian sebelumnya menunjukkan akurasi metode <i>Lexicon Based</i> dan <i>Naive Bayes Classifier</i> , sedangkan hasil penelitian ini menunjukkan perbandingan jumlah ulasan positif dan negatif pada aplikasi Riliv.

2.2 Landasan Teori

Landasan teori dalam penelitian ini sebagai pedoman dan referensi bagi peneliti dalam memahami dan menyelesaikan permasalahan yang diteliti. Landasan teori ini berisi pengetahuan dan informasi yang relevan dengan penelitian, sehingga membantu peneliti untuk memperkuat pemahaman tentang topik penelitian. Berikut ini merupakan landasan teori yang relevan dengan topik penelitian ini.

2.2.1 Analisis Sentimen

Analisis sentimen adalah proses menganalisis dan memahami perasaan atau opini yang terkandung dalam teks. Menurut Brahimi, Touahria dan Tari pada penelitian [14] bahwa tujuan utama analisis sentimen adalah mengolah data teks yang tidak terstruktur menjadi informasi yang memiliki makna. Menurut Rozi, Pramono, dan Dahlan, analisis sentimen berfungsi mengevaluasi pendapat dan kecenderungan pendapat terhadap suatu topik, baik itu bersifat negatif maupun positif. Teknik analisis sentimen memiliki cakupan aplikasi yang luas dan dapat diterapkan di berbagai bidang. Media sosial seperti Twitter dan *platform* distribusi digital seperti *Google Play Store* dan *App Store* menjadi sumber data yang berharga untuk analisis sentimen karena platform ini memungkinkan pengguna untuk secara bebas mengungkapkan pendapat dan emosinya. Dengan menganalisis data teks dari media sosial, peneliti dapat memperoleh wawasan yang berharga tentang opini dan sentimen publik terhadap berbagai topik [14].

Analisis sentimen bertujuan untuk memahami dan mengukur opini, ulasan, atau perasaan yang terkandung dalam teks, seperti ulasan produk, postingan media sosial, atau artikel berita. Melalui analisis sentimen, dapat memperoleh pemahaman tentang perasaan dan sikap orang terhadap suatu topik atau merek. Analisis sentimen menghasilkan informasi untuk membantu perusahaan atau organisasi dalam meningkatkan kepuasan pelanggan, memperbaiki produk atau layanan, serta mengoptimalkan strategi bisnis. Dengan demikian, analisis sentimen memiliki peran penting dalam

pengambilan keputusan yang berdasarkan data dan dapat membantu dalam menghadapi tantangan bisnis yang kompleks [15].

Analisis sentimen merupakan sebuah cabang ilmu komputer yang berfokus pada penemuan dan analisis opini, sentimen, dan emosi yang terkandung dalam teks. Bidang ini mempelajari cara bagaimana komputer dapat memahami dan mengartikan bahasa manusia guna melakukan identifikasi opini dan sentimen yang diekspresikan dalam berbagai bentuk teks, seperti ulasan produk, komentar media sosial, dan artikel berita. Manusia memiliki beragam emosi dasar, seperti rasa takut, kemarahan, jijik, sedih, kaget, dan bahagia. Melalui analisis sentimen dan pemetaan topik dari tanggapan, komentar, atau opini publik, pengguna dapat mengevaluasi sejauh mana kebijakan yang diterapkan telah efisien dan efektif. Dengan demikian, analisis sentimen mengungkapkan tentang bagaimana masyarakat merespons berbagai hal dan situasi [16].

2.2.2 Riliv

Aplikasi Riliv, yang diluncurkan pada tahun 2015 oleh Audrey Maximillian Herli dan Audy Christopher Herl, adalah sebuah *platform* berbasis *mobile* yang bertujuan untuk meningkatkan kesejahteraan psikologis para penggunanya. Aplikasi Riliv adalah salah satu aplikasi yang dikembangkan di Indonesia yang memberikan perhatian khusus pada kesehatan mental pengguna [17].



Gambar 2.1 Aplikasi Riliv

Aplikasi Riliv mengklaim sebagai aplikasi kesehatan mental terbaik di Indonesia. menghadirkan platform komprehensif untuk mengevaluasi, memantau, dan meningkatkan kesehatan mental. Platform ini tidak hanya menawarkan layanan aplikasi, tetapi juga memiliki kehadiran aktif di media sosial, khususnya Instagram, dengan akun bisnis bernama @Riliv. Akun Instagram Riliv telah menarik 292.000 pengikut hingga April 2021, dan setiap postingannya mendapatkan rata-rata 1.000 hingga 25.000 suka. Tim media sosial aplikasi Riliv menunjukkan keahliannya dalam menghasilkan konten visual yang menarik dan informatif dengan menerapkan prinsip desain yang efektif. Pengguna Instagram Riliv tidak hanya menerima konten, tetapi juga terlibat aktif dengan memberikan umpan balik dan komentar pada postingan. Hal ini mendorong aplikasi Riliv untuk memanfaatkan Instagram sebagai media komunikasi dua arah, di mana tim Riliv dapat berinteraksi dengan para pengikutnya dan membangun keterlibatan pelanggan melalui penggunaan konten visual secara berulang [18]

Aplikasi Riliv menawarkan berbagai fitur yang dapat digunakan oleh pengguna, seperti pelacak mood, latihan kebersihan pikiran, layanan konseling, jurnal, perjalanan pribadi, akses ke artikel terkait kesehatan mental, dan video yang informatif [19].

2.2.3 Text Mining

Text mining adalah metode *data mining* guna mengungkapkan informasi yang tersirat dari data teks. Proses *text mining* melibatkan ekstraksi informasi yang tersembunyi dalam data teks dan analisis yang mendalam untuk menghasilkan wawasan yang berbeda dari pengolahan data lainnya. Algoritma dalam *text mining* memiliki kinerja yang serupa dengan *data mining*, tetapi bekerja pada jenis data yang berbeda, yaitu data teks tidak terstruktur. Beberapa proses dalam *text mining* termasuk pengkategorian teks, ekstraksi konsep atau entitas, pengelompokan teks, analisis sentimen, pembuatan taksonomi granular, inferensi dokumen, dan pemodelan hubungan entitas [20].

2.2.4 *Natural Language Processing*

Natural Language Processing (NLP) ialah sebuah teknik dan sub-bidang dalam bidang kecerdasan buatan yang fokus pada pengolahan bahasa alami. NLP dimanfaatkan agar komputer bisa mengerti, memproses, dan mengeluarkan bahasa yang dipakai manusia. Di era digital dan komputasi saat ini, sangat penting untuk mengembangkan mekanisme yang memungkinkan komputer untuk memahami bahasa manusia. NLP membantu mencapai tujuan ini dengan memproses bahasa alami melalui berbagai proses dan algoritme [20].

Tujuan dari *Natural Language Processing* (NLP) adalah untuk memberi kemampuan pada komputer agar bisa menafsirkan, memproses, dan menghasilkan bahasa manusia dengan cara yang menyerupai manusia. Dalam konteks analisis sentimen, NLP memiliki tujuan menganalisis dan memproses ulasan, pendapat, serta emosi dalam teks. Proses ini melibatkan pengelompokan teks ke dalam kategori positif, kategori negatif, atau kategori netral, serta menganalisis ulasan guna mengetahui tanggapan dan persepsi manusia terhadap suatu topik [21].

Sebelum melakukan analisis, penting untuk melakukan normalisasi teks berdasarkan data yang ada. Normalisasi teks dapat dilakukan dengan menggunakan korpus atau kumpulan kata-kata yang tersedia. Contohnya, kata "dgn" dapat dinormalisasi menjadi "dengan". Hal ini penting agar mesin tidak mengalami kesalahan dalam menganalisis makna. Setelah normalisasi, data teks dapat dikategorikan dan diberi label sebagai sentimen positif, negatif, atau netral. Untuk melakukan kategorisasi sentimen, pengguna dapat mengakses sumber daya leksikal yang tersedia. Jika sumber leksikal mengandung banyak data teks yang mengandung kata-kata negatif, maka data teks tersebut dapat dianggap memiliki sentimen negatif. Sebaliknya, jika banyak data teks dalam sumber leksikal yang mengandung kata-kata positif, maka data teks tersebut dapat dianggap memiliki sentimen positif [21].

2.2.5 Naive Bayes

Naive Bayes merupakan metode yang mengandalkan teori probabilitas dalam melakukan klasifikasi. Metode ini tidak memiliki aturan baku dan berfokus pada penggunaan frekuensi atau jumlah kemunculan setiap klasifikasi dalam data pelatihan untuk mendapatkan probabilitas tertinggi. Dalam pengembangan basis data, Naive Bayes melibatkan pembelajaran yang diawasi, dimana sampel data pelatihan memiliki label atau klasifikasi. Metode ini terbagi menjadi dua bagian, yaitu klasifikasi dan regresi. Klasifikasi digunakan ketika variabel harus dikategorikan, misalnya panas atau dingin, sakit atau tidak sakit, dan sejenisnya. Disisi lain, regresi digunakan ketika variabel adalah nilai yang kontinu, seperti bobot atau nilai moneter. Contoh metode lain dalam klasifikasi termasuk *Support Vector Machine* (SVM), *K-Nearest Neighbor* (KNN), dan *Artificial Neural Network* (ANN). Naive Bayes juga menggunakan teknik probabilistik di mana fitur-fitur atau atribut dalam data dianggap saling independen, artinya atribut dalam data dianggap tidak saling bergantung atau mempengaruhi satu sama lain [22].

Naive Bayes suatu metode untuk klasifikasi teks serta data mining dalam analisis sentimen. Algoritma ini memiliki ciri khas dalam membuat asumsi yang kuat mengenai kondisi atau peristiwa yang terjadi. Pendekatan Bayesian pada algoritma ini menggunakan persamaan matematis tertentu [23].

Naive Bayes sebuah metode klasifikasi sederhana dimana setiap atribut dianggap bebas atau independen dan berkontribusi terhadap pengambilan keputusan akhir. Dalam pemrograman, digunakan rumus Bayesian untuk melakukan perhitungan dalam algoritma Naive Bayes sebagai berikut.

$$P(C_i | x) = \frac{P(x | C_i) * P(C_i)}{P(x)} \quad (2.1)$$

Dalam persamaan (2.1) $P(C_i | x)$ merupakan probabilitas kelas C_i jika diberikan fitur x . Sementara itu, $P(x | C_i)$ merupakan probabilitas fitur x muncul di kelas C_i . $P(C_i)$ merupakan jumlah keseluruhan pada kelas C_i . $P(x)$ adalah probabilitas total fitur x [24].

2.2.6 *Crawling*

Teknik *Crawling* adalah teknik untuk mengumpulkan informasi yang tersedia di web. Proses ini dilakukan dengan menjelajahi tautan yang ada di halaman web dan mengambil informasi dari halaman web dan tautan yang dikunjungi. Dengan kata lain, teknik ini memungkinkan pengumpulan data secara otomatis dari berbagai sumber di internet dengan mengikuti dan mengekstrak informasi dari setiap tautan yang diakses [25].

2.2.7 *Preprocessing*

2.2.7.1 *Cleaning*

Data *cleaning* melibatkan langkah – langkah untuk menghapus nilai yang hilang, *noise*, dan data yang tidak konsisten dari sebuah dataset[26].

Data *Cleaning* adalah langkah yang dilakukan untuk mengelola data yang tidak lengkap dalam dokumen. Tujuan dari pembersihan data adalah untuk mengidentifikasi dan menghilangkan data yang ganda, memverifikasi konsistensi data, serta memperbaiki kesalahan yang ada dalam data tersebut. Melalui pembersihan data, data dapat diorganisir dengan baik, relevan, dan terstruktur, serta mencegah kerusakan data [26].

2.2.7.2 *Case Folding*

Case folding adalah pengolahan teks yang fungsinya menyamakan huruf dalam dokumen. Tujuan dari *case folding* adalah untuk memudahkan pencarian teks. Terkadang penggunaan huruf kapital tidak konsisten pada dokumen teks. Untuk itu, penting untuk mempertahankan fitur sensitivitas huruf saat mengonversi semua teks dalam dokumen ke bentuk standar, seperti huruf kecil [27].

2.2.7.3 *Tokenisasi*

Tokenisasi adalah langkah memecah dokumen menjadi unit-unit kecil yang dikenal sebagai token [27]. Tokenisasi juga merupakan tahap

awal dalam pemrosesan data yang menghilangkan tanda baca atau simbol non-huruf [28].

2.2.7.4 Stemming

Stemming adalah langkah mengonversi kata – kata dalam suatu teks menjadi kata dasarnya [28]. *Stemming* membantu mengurangi jumlah kata yang diindeks tanpa mengorbankan makna [29].

2.2.7.5 Stopword Removal

Stopword removal adalah langkah penghapusan kata yang tidak penting dari dokumen. Kata ini biasanya muncul dengan frekuensi tinggi dalam berbagai kalimat, termasuk kata penghubung seperti "adalah", "ke", "di", "dari", serta "dan"[30].

2.2.8 Term Frequency-Inverse Document Frequency (TF-IDF)

Term weighting ialah langkah pembobotan kata dalam analisis sentimen melalui *text mining*. TF-IDF bertujuan untuk mengoptimalkan kemampuan analisis sentimen. Berikut rumus untuk menghitung bobot TF – IDF[31].

$$tf(w,d) = \frac{n(w,d)}{n(d)} \quad (2.2)$$

Pada persamaan (2.2), $tf(w,d)$ digunakan untuk mengevaluasi frekuensi suatu kata dalam sebuah teks atau dokumen, yang kemudian digunakan untuk menentukan seberapa pentingnya kata tersebut dalam konteks tersebut. Sementara $n(w,d)$ merupakan total kata w muncul dalam dokumen d . $n(d)$ merupakan total kata dalam dokumen d [31].

$$idf(w) = \log\left(\frac{N}{n(w)} + 1\right) \quad (2.3)$$

Pada persamaan (2.3), $idf(w)$ digunakan untuk memberikan bobot kepada token berdasarkan seberapa umum atau langka kemunculannya dalam seluruh kumpulan teks. Sementara N merupakan total dokumen. $n(w)$ merupakan jumlah dokumen yang mengandung kata t . Dari persamaan (2.2) dan (2.3) didapat rumus TF – IDF berikut[26].

$$\text{TF-IDF}(w,d) = \text{tf}(w,d) * \text{idf}(w) = \log\left(\frac{N}{n(w)} + 1\right) \quad (2.4)$$

2.2.9 Klasifikasi

Menurut Widodo, klasifikasi adalah teknik yang diterapkan untuk mengelompokkan dataset menjadi kelompok – kelompok berdasarkan kriteria yang telah dipilih [32].

Klasifikasi adalah proses pencarian model atau fungsi untuk mengklasifikasikan objek yang tidak diketahui ke dalam kelas yang tepat. Terdapat dua langkah utama dalam klasifikasi, yaitu langkah pelatihan dan pengujian. Dalam proses pelatihan, digunakan kumpulan data yang telah memiliki label kelas yang diketahui untuk membuat model [33].

2.2.10 WordCloud

WordCloud adalah metode untuk memvisualisasikan data tekstual. *WordCloud* populer untuk penambangan teks karena mudah dimengerti. *WordCloud* digunakan menampilkan gambaran kemunculan kata dalam format yang menarik namun informatif. Besar kecilnya gambar teks pada kata cloud menyesuaikan dengan frekuensi data, semakin sering kata tersebut digunakan maka semakin besar pula kata yang muncul pada kata cloud[34].

2.2.11 Confusion Matrix

Confusion matrix teknik melakukan pengukuran tingkat keakuratan dalam pemrosesan data atau sistem pendukung keputusan. Metode ini mempresentasikan hasil klasifikasi dalam empat kategori yaitu *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN). Untuk menghitung nilai presisi, dapat digunakan persamaan khusus yang telah ditentukan [35].

Rentang akurasi yang diperoleh akan digunakan sebagai pedoman untuk menentukan kategori dengan performa terbaik. Berikut adalah rentang akurasi yang digunakan sebagai pedoman dalam proses penentuan[36].

1. Akurasi 0.90 – 1.00 = *Excellent classification*
2. Akurasi 0.80 – 0.90 = *Good classification*
3. Akurasi 0.70 – 0.80 = *Fair classification*
4. Akurasi 0.60 – 0.70 = *Poor classification*
5. Akurasi 0.50 – 0.60 = *Failure*

		Actual Values		
		Positif (P)	Negatif (N)	Netral (n)
Predict Values	Positif (P)	TP	FP	FP
	Negatif (N)	FN	TN	FN
	Netral (n)	Fn	Fn	Tn

Tabel 2.2 Confusion Matrix[37]

Penjelasan mengenai representasi dari confusion matrix berikut.[38].

- 1 *True Positive* (TP) adalah ketika model berhasil memprediksi dengan benar bahwa sebuah kasus adalah positif. Dalam konteks *confusion matrix*, jika kejadian sebenarnya positif dan prediksi model juga positif, maka disebut sebagai *True Positive*.
- 2 *False Positive* (FP) adalah ketika model memprediksi dengan salah bahwa sebuah kasus adalah positif padahal sebenarnya negatif. Dalam konteks *confusion matrix*, jika kejadian sebenarnya negatif namun prediksi model positif, maka disebut *False Positive*.

- 3 *False Negative* (FN) adalah ketika model memprediksi dengan salah bahwa sebuah kasus adalah negatif padahal sebenarnya positif. Dalam konteks *confusion matrix*, jika kejadian sebenarnya positif namun prediksi model negatif, maka disebut *False Negative*.
- 4 *True Negative* (TN) adalah ketika model berhasil memprediksi dengan benar bahwa sebuah kasus adalah negatif. Dalam konteks *confusion matrix*, jika kejadian sebenarnya negatif dan prediksi model negatif, maka disebut *True Negative*.

Dari keterangan yang terdapat pada Tabel 2.2 di atas, terdapat rumus-rumus yang berguna untuk mengukur performa model, diantaranya sebagai berikut [38].

1. *Accuracy*, mengevaluasi sejauh mana model dapat melakukan klasifikasi dengan tepat. Dalam evaluasi tersebut, jumlah data yang terklasifikasikan dengan benar menjadi penentu akurasi hasil prediksi yang lebih tepat dan akurat. Rumus untuk menghitung *accuracy* dapat dinyatakan menggunakan persamaan (2.2)

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.2)$$

2. *Recall* menggambarkan seberapa baik model dalam mendeteksi kembali informasi yang sebenarnya positif. *Recall* memberikan informasi tentang sejauh mana model dapat mengenali kasus-kasus positif yang benar-benar positif. Dengan kata lain, *recall* mengukur kemampuan model untuk menemukan semua kasus yang sebenarnya positif. Rumus untuk menghitung *recall* dapat dinyatakan menggunakan persamaan (2.3)

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2.3)$$

3. *Precision* menggambarkan tingkat keakuratan prediksi positif dari model. *Precision* dapat dijelaskan sebagai sejauh mana prediksi positif

model benar. Dengan kata lain, *precision* mengukur seberapa sering prediksi positif model sesuai dengan kenyataan. Rumus untuk menghitung *Precision* dapat dinyatakan menggunakan persamaan (2.4)

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2.4)$$

4. *F1-Score* adalah nilai yang mencerminkan keseimbangan antara *precision* dan *recall* yang diukur dengan menggunakan rata – rata harmonik dari kedua matriks tersebut, sehingga *F1-Score* berguna untuk mengukur kinerja keseluruhan model dalam klasifikasi yang akurat. Rumus untuk menghitung *F1-Score* dapat dinyatakan menggunakan persamaan (2.5)

$$\text{F1 Score} = 2 * \frac{(\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})} \quad (2.5)$$

2.2.12 SMOTE (*Synthetic Minority Over-sampling Technique*)

Teknik *Synthetic Minority Oversampling Technique* (SMOTE) adalah algoritma yang sering digunakan untuk menangani masalah data yang tidak seimbang. SMOTE berfungsi dengan menyeimbangkan dataset yang tidak seimbang melalui pembuatan data sintetis baru dari kelas minoritas. Tujuannya adalah untuk meningkatkan efektivitas metode klasifikasi dengan memastikan bahwa semua kelas dalam dataset memiliki representasi yang lebih seimbang.