BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Setelah melakukan kajian mendalam mengenai topik analisis sentimen, ditemukan beberapa jurnal dan penelitian terdahulu yang memiliki relevansi dengan penelitian ini. Penelitian-penelitian tersebut mencakup berbagai aspek dan metodologi yang menjadi landasan bagi penelitian ini. Berikut adalah beberapa penelitian yang relevan yang disajikan dalam Tabel 2.1.

Tabel 2 1 Penelitian Terdahulu

Nama Jurnal	Judul Artikel	Tahun	Nama Peneliti	Hasil Penelitian
e-Proceeding of Management: Vol.8, No.5 Oktober 2021 [8]	Analisis Sentimen Ulasan Aplikasi Spotify Untuk Peningkatan Layanan Menggunakan Algoritma *Naive Bayes*	2021	Mochamad Daffa Rhajendra, Nurvita Trianasari	Dalam penelitian ini, dilakukan analisis sentimen ulasan aplikasi Spotify dengan menggunakan algoritma *Naïve Bayes* dan teknik *e-Servqual* dan mendapatkan *accuracy* model sentimen analisis sebesar 74.85%.
KLIK: Kajian Ilmiah Informatika dan Komputer Vol 4, No 4, Februari 2024 [13]	Analisis Sentimen Aplikasi Tokocrypto Berdasarkan Ulasan Pada Google Play store Menggunakan Metode *Naïve Bayes*	2024	Rizki Adi Saputra, Dion Parisda Ray, Faldy Iriwiensyah	Memahami sentimen ulasan pengguna pada aplikasi Tokocrypto berdasarkan ulasan pengguna aplikasi Tokocrypto. Hasil yang diperoleh menggunakan metode algoritma *Naïve Bayes* yaitu akurasi sebesar 72.22%, *precision* 63.25% dan *recall* 81.40%.

Nama Jurnal	Judul Artikel	Tahun	Nama Peneliti	Hasil Penelitian
MALCOM: Indonesian Journal of *Machine learning* and Computer Science Vol. 4Iss. 2April 2024 [14]	Perbandingan Algoritma SVM dan *Naïve Bayes* dalam Analisis Sentimen Twitter pada Penggunaan Mobil Listrik di Indonesia	2024	Widia Ningsih, Baginda Alfianda, Rahmaddeni, Denok Wulandari	Sentimen diklasifikasikan berdasarkan polaritas teks untuk menentukkan sebagai sentimen positif, negatif, atau netral. Setelah dilakukan perbandingan *Support Vector Machine* memiliki hasil yang lebih baik dibandingkan dengan *Naïve Bayes* dengan akurasi sebanyak 70.82% dan *Naïve Bayes* sebanyak 63.02%
JATI (Jurnal Mahasiswa Teknik Informatika) Vol. 8 No. 1, Februari 2024 [9]	Analisis Sentimen Terhadap Layanan Aplikasi Grab Indonesia Menggunakan Metode *Naïve Bayes*	2024	Ahmad Rifa'I, Risma Ardhani, Denni Pratama, Fatihanursaro	Penelitian ini mengambil permasalahan dari Tingkat kepuasan dan layanan yang diberikan. Klasifikasi ulasan berupa sentimen positif dan negatif yang dilakukan menggunakan metode *Naïve Bayes* dengan hasil akurasi sebesar 84.36%.

Nama Jurnal	Judul Artikel	Tahun	Nama Peneliti	Hasil Penelitian
Ultima Infosys : Jurnal Ilmu Sistem Informasi, Vol. 14, No. 2 | December 2023 [10]	Sentimen Analysis of *User* Satisfaction Towards Sales Promotion of Gojek Application Service Using SVM	2023	Calandra Alencia Haryani, Gabrielle Florencia, Andree E. Widjaja, Hery, Ferry V. Ferdinand	Penelitian ini bertujuan untuk memahami sentimen pelanggan terhadap promosi atau penawaran untuk mengembangkan strategi promosi yang lebih baik dengan meningkatkan kepuasan pelanggan. Hasil penelitian memperoleh akurasi sebesar 93% dan dilengkapi visualiasi untuk mempresentasikan frekuensi *Wordcloud* secara keseluruhan
Scientica Jurnal Ilmiah Sain dan Teknologi (2023),1(2): 110–120 [15]	Analisis Sentimen Menggunakan Metode Naive Bayesdan *Support Vector Machine*pada Ulasan Aplikasi Joox Music	2023	Wayan Pondra Lesmana, Andri Wijaya	Hasil penelitian ini memberikan masukan-masukan informasi kepada pengembang Joox Music untuk memahami opini dari pengguna serta merespon masukan pelanggan dengan lebih efektif.
STORAGE – Jurnal Ilmiah Teknik dan Ilmu Komputer Vol. 2No. 3, Agustus 2023 [16]	Analisis Sentimen Review Aplikasi WETV pada Platform Twitter Menggunakan *Support Vector Machine*	2023	Vina Alviani, Syariful Alam, Imay Kurniawan	Penelitian ini bertujuan untuk memberikan masukan untuk WETv dalam perbaikan terhadap aplikasi untuk mengatasi permasalahan yang ada. Jumlah data yang dipakai berjumlah 4024 data dan nilai akurasi SVM memiliki nilai 89%

Nama Jurnal	Judul Artikel	Tahun	Nama Peneliti	Hasil Penelitian
KLIK: Kajian Ilmiah Informatika dan Komputer Vol 3, No 6, Juni2023 [17]	Analisis Sentimen Ulasan Pelanggan Pada Aplikasi Fore CoffeeMenggunakan Metode *Naïve Bayes*	2023	Tia Anggita Sari, Estu Sinduningrum, Firman noor Hasan	Penelitian ini menghasilkan sebuah klasifikasi kepuasan pelanggan dengan menggunakan Algoritma *Naïve Bayes*. Nilai akurasi yang didapat setelah diolah diperoleh presentasi sebesar 74.28%
Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi) Vol. 7No. 3(2023) [18]	Sentiment Analysis of Public Acceptance of Covid-19 Vaccines Types in Indonesia using *Naïve Bayes*, *Support Vector Machine*, and Long Short-Term Memory (LSTM)	2023	Dinar Ajeng Kristiyanti, Sri Hardani	Hasil penelitian bertujuan untuk melakukan analisis sentiment terhadap penerimaan public terhadap jenis vaksin Covid-19 yang digunakan di Indonesia menggunakan data Twitter dan membandingkan algoritma NB, SVM, dan LTSM. Hasil penelitian menunjukkan skor SVM merupakan model terbaik sebesar 84,99%, *Naïve Bayes* sebesar 84,65% dan LTSM sebesar 82,97%
JIKA (Jurnal Informatika) Universitas Muhammadiyah Tangerang Vol 7 No 1, [19]	Sentiment Analysis Public Opinion of CFW (Citayam Fashion Week) on Media Social Twitter using *Naïve Bayes* Classifier	2023	Angga Aditya Permana, Permana Perdana Putra	Penelitian ini menganalisa sentimen pandangan masyarakat yang berbeda ada yang menyikapi dengan positif ataupun mengkritik maka perlu dilakukan analisa sentimen pada media sosial twitter. Hasil evaluasi menggunakan *Naïve Bayes* sebesar 84%

Pada tabel 2.1 terdapat beberapa penelitian terdahulu. Jurnal pertama yang ditulis oleh Mochamad Daffa Rhajendra dan Nurvita dengan judul "Analisis Sentimen Ulasan Aplikasi Spotify Untuk Peningkatan Layanan Menggunakan Algoritma *Naive Bayes*" menganalisis 3600 ulasan Spotify dari Google Play dengan algoritma Naive Bayes dan metode e-Servqual. Penelitian ini fokus pada kualitas layanan seperti *app design, reliability, responsiveness, trust*, dan *personalization*. Kelebihannya termasuk menggunakan dimensi e-Servqual untuk mengevaluasi kualitas layanan, serta memberikan rekomendasi perbaikan layanan yang konkret. Namun, kekurangannya adalah kurangnya penjelasan detail mengenai aspek bisnis, yang dapat membatasi pemahaman yang lebih luas tentang faktor-faktor yang memengaruhi kepuasan pengguna [8].

Jurnal kedua yang ditulis oleh Rizki Adi Saputra, Dion Parisda Ray, dan Faldy Iriwiensyah dengan judul "Analisis Sentimen Aplikasi Tokocrypto Berdasarkan Ulasan Pada Google Play store Menggunakan Metode *Naïve Bayes*" Penelitian ini menggambarkan langkah-langkah yang terstruktur, mulai dari pengumpulan data hingga evaluasi, memberikan gambaran yang komprehensif tentang proses analisis sentimen. Namun, kekurangannya terletak pada kurangnya informasi tentang faktor-faktor penentu dalam proses klasifikasi data menggunakan algoritma *Naive Bayes*[13].

Jurnal ketiga ditulis oleh Widia Ningsih, Baginda Alfianda, Rahmaddeni, dan Denok Wulandari dengan judul "Perbandingan Algoritma SVM dan *Naïve Bayes* dalam Analisis Sentimen Twitter pada Penggunaan Mobil Listrik di Indonesia". Kelebihan dari penelitian ini adalah penggunaan dua algoritma klasifikasi yang berbeda, yaitu *Naive Bayes* dan *Support Vector Machine*, memberikan perbandingan yang komprehensif tentang performa klasifikasi, sehingga memungkinkan pemilihan metode yang lebih efektif. Namun, kekurangannya adalah kurangnya rincian tentang teknik preprocessing yang digunakan dan proses pembangunan model klasifikasi [14].

Jurnal keempat ditulis oleh Ahmad Rifa'I, Risma Ardhani, Denni Pratama, dan Fatihanursaro dengan judul "Analisis Sentimen Terhadap Layanan Aplikasi

Grab Indonesia Menggunakan Metode *Naïve Bayes*". Kelebihan penelitian ini terletak pada penerapan analisis sentimen terhadap layanan aplikasi Grab, yang memberikan pemahaman mendalam tentang tingkat kepuasan pengguna berdasarkan ulasan yang diberikan. Penggunaan metode klasifikasi Naive Bayes yang menghasilkan tingkat akurasi yang tinggi, dengan nilai *precision* mencapai 99,64%, *recall* sebesar 87,12%, dan tingkat akurasi sebesar 84,36%. Namun, kekurangan yang mungkin ada adalah kurangnya detail tentang metodologi pengumpulan data dan teknik preprocessing yang digunakan. Meskipun demikian, adopsi algoritma *Naive Bayes* yang menghasilkan tingkat akurasi yang tinggi menjadi salah satu alasan dalam analisis sentimen aplikasi [9].

Jurnal kelima ditulis oleh Calandra Alencia Haryani, Gabrielle Florencia, Andree E. Widjaja, Hery, dan Ferry V. Ferdinand dengan judul "*Sentimen Analysis of User Satisfaction Towards Sales Promotion of Gojek Application Service Using SVM*". Penelitian ini menyoroti pentingnya kepuasan pelanggan dalam menjaga kelangsungan sebuah perusahaan, terutama perusahaan teknologi seperti Gojek. Meskipun perusahaan telah menerapkan strategi promosi atau penawaran untuk meningkatkan daya tarik dan kepuasan pelanggan, analisis sentimen menunjukkan bahwa masih banyak keluhan yang mengindikasikan kebutuhan untuk meningkatkan kepuasan pelanggan. Kelebihan dari penelitian ini adalah memberikan wawasan yang mendalam tentang persepsi pelanggan terhadap promosi atau penawaran yang diberikan oleh perusahaan Gojek. Penelitian ini juga memberikan rekomendasi untuk meningkatkan strategi promosi guna meningkatkan kepuasan pelanggan. Namun, kekurangannya terletak pada kurangnya pembahasan mengenai metodologi yang digunakan dalam analisis sentimen, seperti metode preprocessing dan penggunaan algoritma SVM, sehingga membatasi pemahaman yang lebih mendalam tentang proses analisis[10].

Jurnal keenam ditulis oleh Wayan Pondra Lesmana dan Andri Wijaya dengan judul "Analisis Sentimen Menggunakan Metode Naive Bayes dan Support Vector Machine pada Ulasan Aplikasi Joox Music". Kelebihan dari penelitian ini terletak pada hasil analisis sentimen memberikan wawasan berharga bagi pengembang untuk memahami opini pengguna dan merespons masukan

pelanggan secara lebih efektif, yang dapat meningkatkan pengalaman pengguna dalam menggunakan aplikasi Joox Music. Namun, kekurangannya adalah kurangnya detail tentang metode pengumpulan data ulasan pengguna, seperti proses manual pengumpulan dan cara penyeleksian evaluasi lebih lanjut terhadap hasil analisis[15].

Jurnal ketujuh ditulis oleh Vina Alviani, Syariful Alam, dan Imay Kurniawan dengan judul "Analisis Sentimen Review Aplikasi WETV pada Platform Twitter Menggunakan Support Vector Machine". Kelebihan dari penelitian ini adalah memberikan gambaran yang cukup komprehensif tentang hasil penelitian analisis sentimen pengguna terhadap aplikasi WeTV dengan menyajikan detail proses metodologi, evaluasi, sampai dengan kesimpulan. Namun, kekurangannya terletak pada pengulangan dalam menyampaikan informasi yang sama dalam beberapa bagian dan kurang memberikan tambahan insight yang signifikan. Pemilihan algoritma Support Vector Machine (SVM) didasarkan pada akurasinya yang tinggi, mengindikasikan kehandalan SVM dalam mengolah data sentimen secara efektif, yang mempengaruhi keputusan penggunaan algoritma tersebut dalam analisis sentimen[16[.

Jurnal kedelapan ditulis oleh Tia Anggita Sari, Estu Sinduningrum, dan Firman Noor Hasan dengan judul "Analisis Sentimen Ulasan Pelanggan Pada Aplikasi Fore Coffee Menggunakan Metode Naïve Bayes". Kelebihannya terdapat dalam kebutuhan untuk menyajikan lebih banyak detail tentang metodologi yang digunakan dalam pengumpulan data, seperti teknik web scraping yang digunakan serta sumber data yang spesifik, dan cara pengelompokkan data yang dilakukan penjabaran tentang bagaimana hasil tersebut dapat memberikan wawasan yang berguna bagi Fore Coffee dalam meningkatkan pelayanan atau produk mereka bisa lebih diperinci. Namun, kekurangannya terletak pada penekanan pada akurasi tinggi dari algoritma Naïve Bayes tanpa memberikan konteks atau pembanding dengan metode lain dapat membatasi pemahaman[17].

Jurnal kesembilan ditulis oleh Dinar Ajeng Kristiyanti dan Sri Hardani dengan judul "Sentiment Analysis of Public Acceptance of Covid-19 Vaccines Types in Indonesia using Naïve Bayes, Support Vector Machine, and Long Short-Term Memory (LSTM)". Kelebihan dari penelitian ini adalah bahwa itu

memberikan gambaran yang jelas tentang metodologi penelitian yang dilakukan untuk menganalisis sentimen masyarakat terhadap vaksin Covid-19 di Indonesia Namun, ada beberapa kekurangan seperti penelitian ini hanya menggunakan data dari satu platform media sosial, yaitu Twitter, yang mungkin tidak sepenuhnya mencerminkan keragaman opini masyarakat secara menyeluruh.menggunakan data dari Twitter. Penelitian tersebut menggunakan berbagai algoritma seperti Naïve Bayes, SVM, dan LSTM untuk mencapai pemodelan yang akurat[18].

Jurnal kesepuluh ditulis oleh Angga Aditya Permana dan Permana Perdana Putra dengan judul "Sentiment Analysis Public Opinion of CFW (Citayam Fashion Week) on Media Social Twitter using Naïve Bayes Classifier". Kelebihan dari teks di atas adalah memberikan gambaran yang jelas tentang fenomena Citayam Fashion Week (CFW) dan dampaknya terhadap minat remaja dalam industri fashion lokal. Teks tersebut menggambarkan adanya perbedaan pandangan di masyarakat terhadap CFW. Hasil penelitian menunjukkan keefektifan metode tersebut dalam mengklasifikasikan sentimen terhadap acara tersebut, dengan tingkat akurasi yang signifikan. Namun, ada beberapa kekurangan seperti teks tidak memberikan detail yang cukup tentang metodologi penelitian yang digunakan, seperti proses pengambilan data, pre-processing, dan klasifikasi data. Informasi yang lebih rinci tentang langkah-langkah penelitian ini[19].

Kebaruan dari penelitian ini terletak pada perbandingan antara dua model algoritma, yaitu *Support Vector Machine* dan *Naive Bayes*, untuk menentukan model yang lebih baik dengan harapan hasil dapat memberikan pemahaman yang lebih dalam tentang keunggulan SVM dan penggunaan metodologi CRISP-DM dalam konteks analisis data. SVM memiliki sejumlah keunggulan, seperti kemampuan generalisasi yang baik, serta efisiensi dalam memproses data kompleks, yang membuatnya lebih unggul daripada *Naive Bayes* dalam banyak situasi, terutama pada dataset dengan dimensi tinggi dan kompleksitas yang tinggi. Sementara itu, kebaruan lain dalam penelitian ini adalah pencapaian tingkat akurasi yang melampaui jurnal acuan yang digunakan, serta penerapan pendekatan penelitian yang lebih terstruktur dan komprehensif, seperti metode CRISP-DM, yang memastikan setiap langkah dalam proses analisis data

dilakukan dengan cermat dan teliti, sehingga hasilnya menjadi lebih akurat.

Secara keseluruhan, penelitian terdahulu ini menunjukkan bahwa baik *Naïve Bayes* maupun SVM dapat menunjukkan kinerja yang unggul tergantung pada konteks tertentu. Oleh karena itu, penelitian ini bertujuan untuk memberikan pemahaman yang lebih mendalam melalui perbandingan langsung antara kedua algoritma tersebut.

2.2 Teori tentang Topik Skripsi

2.2.1 Spotify

Spotify adalah sebuah platform *streaming* musik digital yang berasal dari Swedia dan didirikan pada tahun 2008. Platform ini telah menjadi pusat perhatian bagi para pecinta musik di seluruh dunia. Dengan berbagai lagu dari penyanyi-penyanyi terkenal, Spotify terus memperluas jangkauan dan pendengarnya. Salah satu keunggulan Spotify adalah memberikan peluang bagi para penyanyi untuk menyebarkan karya-karya mereka melalui platform ini. Masyarakat dapat dengan mudah mendengarkan berbagai jenis lagu di mana saja dan kapan saja, membuat pengalaman mendengarkan musik menjadi lebih fleksibel. Spotify menyediakan dua opsi layanan, yakni versi gratis dan peningkatan ke Spotify *Premium*. Dengan berlangganan Spotify *Premium*, pengguna dapat menikmati fitur eksklusif seperti mendengarkanmusik tanpa iklan[20].

2.2.2 Analisis Sentimen

Analisis sentimen adalah proses memahami, mengekstraksi informasi, dan mengolah data teks untuk mengidentifikasi sentimen positif atau negatif. Pengguna internet mempunyai hak menyatakan pendapat dan

memberikan *rating* maupun ulasan mengenai pendapat pribadi. Dalam bisnis, analisis sentimen banyak digunakan untuk kebutuhan bisnis mereka dalam mendeteksi reputasi merek dan memahami pelanggan. Sentimen positif bisa dijelaskan dengan menggunakan kata-kata seperti "istimewa", "fantastis", "hebat". Sedangkan sentiment negatif dapat dinyatakan dengan menggunakan kata-kata seperti "jelek", "menyedihkan", "mengecewakan" [21].

Analisis sentimen memiliki beberapa keuntungan yaitu mendapatkan informasi dan wawasan sehingga dapat membantu untuk meningkatkan produk dan layanan, mengetahui pemahaman yang lebih baik tentang pendapat pelanggan, mengetahui ulasan positif dan negatif lebih cepat. Dengan memanfaatkan alat analisis sentimen, kita dapat mengidentifikasi ulasan yang menggambarkan keberhasilan dan kesuksesan sebagai bagian dari kategori positif. Ulasan negatif dapat diindikasikan sebagai ulasan buruk masalah dengan produk atau layanan[22].

Tujuan dari analisa adalah untuk secara otomatis mengidentifikasi dan mengklasifikasikan pendapat pengguna yang telah dituangkan dalam bentuk teks. Selain itu, dalam beberapa kasus analisis sentimen dapat *customer support* menjadi lebih responsif dan perhatian terhadap keluhan pengguna. Dengan bantuan analisis sentimen, kita dapat memahami respon maupun opini masyarakat terhadap topik tertentu, sehingga dapat lebih mudah dalam memahami hasil akhir sentimen dari keseluruhan konteks topik

2.2.3 *CountVectorizer*

CountVectorizer merupakan sebuah alat pemrosesan bahasa alami (NLP) untuk menghasilkan vektor kalimat. Alat ini digunakan untuk menerjemahkan teks ke dalam vektor dengan cara menghitung berapa kali frekuensi setiap kata muncul dalam teks. Dengan menggunakan *CountVectorizer*, teks dapat diubah menjadi representasi vektor yang dapat digunakan dalam model pembelajaran mesin untuk melakukan berbagai

tugas seperti analisis sentimen, klasifikasi dokumen, atau tugas klasifikasi teks lainnya[23].

2.2.4 *Confusion Matrix*

Konsep *Confusion Matrix* melakukan akurasi data dengan perhitungan dari *Data Mining* maupun sistem pendukung sebuah keputusan. Terdapat 4 istilah sebagai representasi hasil proses klasifikasi. Keempat istilah tersebut antara lain *True positive* (TP), *True negative* (TN), *False positive* (FP), dan *False Negative* (FN)[24].

True positive (TP) : data memiliki nilai positif dan terdeteksi benar bahwa data positif

False positive (FP) : data memiliki nilai negatif tetapi terdeteksi sebagai data positif. *False Negative* (FN) : data memiliki nilai positif tetapi terdeteksi sebagai data negatif.

True negative (TN) : data memiliki nilai negatif dan terdeteksi benar sebagai data negatif.

Nilai dari *Confusion Matrix*, akan menghasilkan nilai *accuracy*, *precision*, *recall*, dan *F1-Score* dengan persamaan pada tabel 2.2 sebagai berikut:

Tabel 2. 2 *Confusion Matrix* [24]

	True	False
True (*Positive*)	TP (*True positive*)	FP (*False positive*)
False (*Negative*)	FN (*False Negative*)	TN (*True negative*)

2.2.4.1 Accuracy

Accuracy memprediksi bahwa seberapa akurat dalam mengklasifikasikan model dengan benar. *Accuracy* merupakan rasio

untuk memprediksi positif dan negatif dari seluruh isi data. Berikut pada rumus 2.1 adalah perhitungan nilai *accuracy*[24] :

$$Accuracy = \frac{TN+TP}{TN+FP+FN+TP}$$

Rumus 2. 1 Perhitungan Nilai *Accuracy*

2.2.4.2 Recall

Recall mengambil keunggulan model dalam mengambil sebuah informasi. *Recall* mempresentasikan persentase prediksi hasil *True positive* dengan jumlah keseluruhan *data positive*. Berikut pada rumus 2.2 merupakan rumus perhitungan nilai *Recall*[24] :

$$Recall = \frac{TP}{TP+FN}$$

Rumus 2. 2 Perhitungan Nilai *Recall*

2.2.4.3 Precision

Precision menghasilkan akurasi data yang diminta dan hasil prediksi yang telah disediakan oleh model. *Precision* disebut juga sebagai rasio prediksi positif. Berikut pada rumus 2.3 merupakan rumus perhitungan nilai *precision*[24] :

$$Precision = \frac{TP}{TP+FP}$$

Rumus 2. 3 Perhitungan Nilai *Precision*

2.2.4.4. F1-Score

F1-Score menggambarkan perbandingan antara rata-rata presisi dan *recall*. *Score* memprediksi hasil positif palsu dan negatif palsu. Berikut pada rumus 2.4 merupakan rumus perhitungan *F1-Score*[24] :

$$\text{FI} - score = F1 = score = 2 \times \frac{precision \times recall}{precision+recall}$$

Rumus 2. 4 Perhitungan Nilai *F1-Score*

2.3. Teori tentang *Framework* / Algoritma yang digunakan

2.3.1 CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISP-DM) adalah suatu kerangka kerja proses yang digunakan dalam *Data Mining dan* analisis data guna membantu perusahaan memahami serta mengatasi tantangan bisnis. mereka secara sistematis. *CRISP-DM* menyediakan suatu kerangka kerja terstruktur untuk mengelola proyek *Data Mining* dari tahap awal hingga implementasi solusi, memberikan pendekatan yang terorganisir untuk menyelesaikan tantangan bisnis[25]. Metodologi ini terdiri dari enam tahapan, yaitu *Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation,* dan *Deployment*. Berikut pada gambar 2.1 dibawah merupakan tahapan *CRISP-DM*:

Gambar 2 1 Tahapan *CRISP-DM* [25]

a. *Business Understanding* (Pemahaman Bisnis)

Tahap *Business Understanding*, fokusnya adalah memahami kebutuhan dan tujuan dari perspektif bisnis, yang kemudian diterjemahkan menjadi definisi masalah dalam konteks *Data Mining*. Langkah selanjutnya adalah merumuskan rencana dan strategi untuk mencapai tujuan *Data*

Mining yang telah ditetapkan, dengan mempertimbangkan informasi dan pengetahuan yang telah dikumpulkan.

b. *Data Understanding* (Pemahaman Data)

Data Understanding berkaitan dengan memperoleh pemahaman yang komprehensif tentang data yang tersedia untuk proyek *Data Mining*, termasuk pengumpulan data, deskripsi data, eksplorasi pola data, dan evaluasi kualitas data.

c. *Data Preparation* (Persiapan Data)

Dalam tahap *Data Preparation* dalam *CRISP-DM*, fokusnya adalah membangun *dataset* akhir dari data mentah. Proses ini melibatkan beberapa langkah, termasuk pembersihan data untuk mengatasi nilai-nilai yang tidak valid atau hilang, pemilihan data untuk menentukan *record* dan atribut-atribut yang relevan, serta melakukan transformasi data agar sesuai dengan kebutuhan analisis. Hasil dari tahap ini akan digunakan sebagai masukan dalam tahap pemodelan untuk proses *Data Mining* selanjutnya.

d. *Modelling* (Pemodelan)

Tahap Data *Modeling* merupakan langkah keempat setelah *Data Preparation*, di mana model atau teknik analisis diterapkan untuk mengekstraksi informasi berharga dari data yang telah dipersiapkan sebelumnya. Ini melibatkan penerapan berbagai teknik *Data Mining* seperti pemodelan statistik atau pembelajaran mesin, sesuai dengan tujuan proyek dan jenis data yang digunakan. Model yang dihasilkan digunakan untuk mengidentifikasi pola atau tren, membuat prediksi, atau klasifikasi, serta mendukung pengambilan keputusan, sering kali melibatkan eksperimen dengan berbagai model untuk menemukan yang paling sesuai dengan data dan tujuan proyek.

e. *Evaluation* (Pengujian)

Evaluasi dalam *CRISP-DM* melibatkan penilaian model atau teknik analisis untuk memverifikasi kualitas dan efektivitasnya. Proses evaluasi menggunakan metrik kinerja untuk mengidentifikasi kelemahan atau keunggulan dalam model, sehingga dapat ditingkatkan jika perlu, dengan tujuan memastikan bahwa solusi yang dihasilkan memberikan nilai bagi bisnis atau organisasi yang terkait.

f. Deployment

Setelah model dibuat, diuji, dan dievaluasi menggunakan data validasi, tahap *deployment* dapat dilakukan dengan cara menyusun laporan.

2.3.2 Naïve Bayes

Teori Thomas Bayes mengungkapkan bahwa memprediksi dari pengalaman sebelumnya untuk memprediksi peluang masa depan. Pernyataan tersebut digabungkan dengan pendekatan Naive yang mengasumsikan bahwa atribut-atribut tidak memiliki ketergantungan satu sama lain. Salah satu ciri utama *Naïve Bayes Classifier* yaitu memiliki asumsi independensi cukup kuat dari setiap kondisi atau peristiwa.

Kelebihan menggunakan algoritma *Naïve Bayes* diantaranya (1) algoritma dapat digunakan untuk multi-kelas maupun biner dalam masalah klasifikasi keduanya , (2). Pengkategorian dokumen dapat dipersonalisasi sesuai dengan tujuan, (3). Klasifikasi *Naïve Bayes* mudah diimplementasikan dan cepat, (4). Dapat digunakan untuk data kuantitatif maupun kualitatif, (5). Data tidak diharuskan memiliki jumlah data yang banyak. Berikut rumus 2.5 merupakan rumus teorema *Naïve Bayes*[26]:

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

Rumus 2 5 *Naïve Bayes*

Berikut merupakan penjelasan rumus pada gambar 2.5 :

Keterangan :

X = Data dengan *class* yang belum diketahui

H = Hipotesis data X merupakan suatu *class* spesifik

P(H|X) = Probabilitas hipotesis H berdasarkan kondisi x

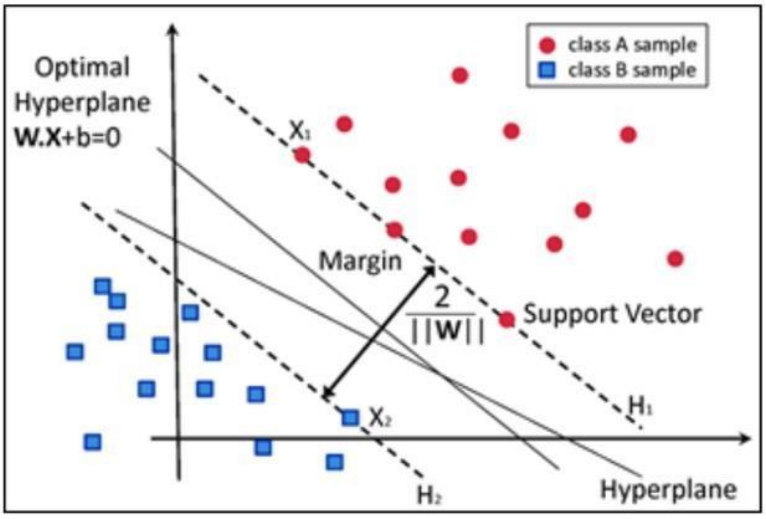
P(H) = Probabilitas hipotesis H

P(X|H) = Probabilitas X berdasarkan kondisi tersebut

P(X) = Probabilitas dari X

2.3.3 Support Vector Machine

Support Vector Machine adalah metode dalam pengklasifikasi sederhana serta intuitif yang disebut dengan pengklasifikasi margin maksimal. Algoritma SVM termasuk dalam kategori *Supervised Learning*, yang berarti data yang digunakan *machine learning* merupakan data yang telah memiliki label sebelumnya. Ketika saat proses penentuan keputusan, mesin akan mengkategorikan data *testing* ke dalam label sesuai dengan karakteristik. Berikut gambar 2.2 dibawah merupakan contoh *Hyperlane* SVM:



Gambar 2 2 Hyperlane SVM [27]

Algoritma SVM bekerja dengan menetapkan batas antara dua kelas yang memiliki jarak maksimum dari data terdekat. Untuk mendapatkan

batas, diperlukan mencari *hiperplan* terbaik dalam ruang *input* dengan mengukur marginnya dan mencari titik maksimalnya[27]. Konsep *kernel* digunakan untuk pemecahan masalah secara n*online*ar di ruang kerja berdimensi tinggi[28], untuk mencari *hyperplane* maksimal margin antar kelas data. *Hyperplane* berguna untuk membedakan dua grup yaitu +1 dan *class* -1 dimana pada setiap kelas memiliki *pattern* sendiri.

2.3.4 Text Mining

Tahap ini dilakukan pengumpulan teks dari dokumen untuk *Text Mining* adalah metode pengumpulan text dari dokumen untuk mendapatkan hasil sesuai dengan tujuan. Proses *Data Mining* dapat diperoleh dari beberapa sumber salah satunya web. *Text Mining* merupakan bagian dari *Data Mining* yang melibatkan analisis kumpulan dokumen dengan menggunakan alat atau metode kategorisasi dalam *Data Mining*[29]. Tujuan dari *Text Mining* dalam analisis sentimen adalah untuk mengidentifikasi dan menganalisis sentimen positif, negatif, atau netral yang terkandung dalam teks seperti pendapat, kritik, ulasan, dan *review*. Apabila *Text Mining* dikelola dengan benar, hasil akhir *Text Mining* dapat memberikan informasi berguna yang dapat membantu individu maupun organisasi membuat sebuah keputusan bagi sebuah bisnis.

2.3.4.1 Data preprocessing

Data preprocessing adalah implementasi dari *Text Mining*. Text *preprocessing* adalah metode mempersiapkan teks yang tidak terstruktur menjadi pemilihan data teks agar lebih terstruktur dan siap untuk diolah melalui serangkaian langkah seperti *case folding*, *tokenizing*, *stemming* dan *stopward removal*[30].

2.2.4.2. *Case folding*

Case folding adalah sebuah proses dalam pengolahan teks yang secara seragam mengubah semua huruf besar menjadi huruf kecil, dan menghapus karakter non-alfanumerik dan tanda baca.

Dengan menerapkan *case folding*, teks bisa diatur sedemikian rupa sehingga memudahkan dalam perbandingan, pencarian, dan analisis.

Case folding mengubah penggunaan semua huruf kapital menjadi huruf kecil. Huruf dan angka, tanda baca, dan karakter lain yang tidak mengandung spasi juga dihilangkan. *Case folding* mengubah semua huruf kecil. Misalnya, "Musiknya Bikin Ngantuk" maka setelah dilakukan *case folding* akan berubah menjadi "musiknya bikin ngantuk".

2.3.4.3 Tokenizing

Pada proses tokenisasi, teks dipecah menjadi unit-unit yang lebih kecil yang disebut token. Tahap ini juga melibatkan penghapusan tanda baca, angka, dan karakter lain yang tidak mempengaruhi pemrosesan kata. Proses ini bertujuan untuk mempersiapkan teks agar lebih mudah diproses oleh model pemrosesan bahasa alami atau algoritma pemrosesan teks lainnya. Setelah teks dipisahkan menjadi token, analisis lebih lanjut seperti pemrosesan kata atau pengkodean dapat dilakukan

2.3.4.4 Stopword Removal

Pada tahap ini kata diambil hanya kata penting saja dari hasil token dengan menghapus kata yang kurang penting. Terdapat dua metode dalam *Stopword Removal*, yakni :

1. Stoplist

Stoplist adalah daftar kata-kata yang dianggap kurang penting atau tidak relevan dalam konteks tertentu, seperti analisis teks atau pemrosesan bahasa alami. Tujuannya adalah untuk mengidentifikasi kata-kata tersebut dalam kumpulan data dan menghapusnya agar tidak mempengaruhi tahapan analisis atau pemrosesan berikutnya. Dengan menghilangkan kata-kata yang dianggap tidak penting,

proses ini dapat membantu meningkatkan efisiensi dan akurasi dari analisis data selanjutnya.

2. Wordlist

Wordlist adalah daftar kata-kata yang dianggap penting atau relevan dalam suatu konteks tertentu. Proses penyusunan *wordlist* melibatkan identifikasi kata-kata kunci yang akan dipertahankan dan digunakan dalam tahapan analisis atau pemrosesan berikutnya. Kata-kata yang masuk dalam *wordlist* dipilih sesuai kebutuhan aplikasi atau penelitian serta tujuan analisis yang akan dilakukan. Sebaliknya, kata-kata yang tidak termasuk dalam *wordlist* akan dihapus atau diabaikan pada proses selanjutnya. Hal ini memastikan bahwa hanya informasi yang relevan yang dipertahankan untuk pengolahan lebih lanjut. Dengan menggunakan *wordlist*, proses pemrosesan data dapat lebih terfokus pada informasi yang penting dan relevan[30].

2.3.4.5 Stemming

Pada tahap ini bentuk dasar kata diubah menjadi sesuai dengan kaidah Bahasa Indonesia. *Stemming* dilakukan menggunakan *library* yang terdapat di Python yaitu Sastrawi. *Stemming* membantu merubah kata-kata ke bentuk dasarnya dengan menghilangkan afiks-afiks seperti awalan dan akhiran yang memungkinkan analisis teks yang lebih efektif. Hal ini berguna dalam berbagai aplikasi seperti indeksasi, pencarian, dan analisis teks.

2.4 Teori tentang *Tools* / *Software* yang digunakan

2.4.1 Google Play Store

Google Play Store adalah aplikasi resmi yang diciptakan oleh Google untuk suatu perangkat baik web maupun android. Google Play telah diinstal sebelumnya secara *default* di berbagai media elektronik seperti handphone, tablet, Android, Android TV, dan perangkat Google TV

lainnya. Melalui *browser* dan web Google Play Store menyediakan layanan dapat diakses secara *online*[30]. Google Play Store menyediakan fitur agar para pengguna dapat mengisi di kolom ulasan dan *rating* berdasarkan opini masing-masing individu. Berdasarkan *feedback* pengguna dapat menjadikan suatu gambaran umum dan sebagai bahan penelitian.

Google Play store merupakan wadah untuk mengunduh ataupun membeli produk digitalGoogle. Google Play store memiliki beberapa jenis layanan diantaranya Google Play Books, Google Play Film & TV, Musik, Google Play Apps dan*Games*. Beragam fitur disediakan oleh Google Play store seperti terdapat fitur metode pembayaran, informasi hiburan terkini, *Subscription*, dan *Wishlist*. Salah satu tujuan Utama Play store adalah menyediakan wadah bagi para pengguna untuk dapat mengunduh aplikasi dan game[29].

2.4.2 Python

Python adalah bahasa pemrograman perancangan yang berfokus pada membaca sebuah kode. Bahasa Python menggabungkan kapabilitas, kemampuan, dan sintaks kode dengan sangat jelas. Python umumnya digunakan sebagai bahasa *script* dan cakupan penggunaannya sangat luas dapat digunakan untuk berbagai program untuk pengembangan serta dapat digunakan pada berbagai platform sistem operasi[30].

Python merupakan bahasa pemrograman yang seringkali dipilih untuk kegiatan analisis data karena fleksibilitas, kekuatan, dan beragamnya perpustakaan yang tersedia, seperti*Pandas*, *NumPy*, dan *Matplotlib*. Dalam konteks analisis data, Python memberikan kemampuan bagi para analis untuk menyelesaikan berbagai tugas, mulai dari pengolahan data hingga visualisasi dengan mudah. Karakteristik kode Python yang mudah dibaca dan ditulis, menjadikannya pilihan yang populer di kalangan praktisi analisis data dari beragam latar belakang. Selain itu, kemampuan Python dalam mengelola berbagai jenis data, termasuk data terstruktur dan tidak terstruktur, menjadikannya pilihanutama dalam pengolahan dan eksplorasi data secara efisien.

2.4.3 Google Colab

Google Colab adalah layanan komputasi *cloud* yang disediakan oleh Google. Platform ini menyediakan berbagai pustaka *machine learning* seperti *NumPy, TensorFlow*, *Pandas*, *Matplotlib*, dan lainnya. Kelebihan utama Colab adalah kemudahannya dalam membuat visualisasi data tanpa perlu menginstal perangkat lunak tambahan di komputer pengguna. Selain itu, Google Colab juga digunakan sebagai software untuk melakukan scrapping dan pemrosesan data dan melakukan operasi pengolahan data secara langsung. Colab sangat sesuai untuk pengembangan dan penelitian di bidang ilmu data, kecerdasan buatan, dan pemrosesan bahasa alami.

Selain itu, Google Colab mendukung berbagai format *file*, termasuk *notebook Jupyter*, yang memudahkan pengguna dalam membuat, menjalankan, dan berbagi *notebook*. Fitur ini juga memungkinkan kolaborasi *online* yang mudah bagi pengguna yang bekerja dalam tim[31