# BAB II
# LANDASAN TEORI

## 2.1 Penelitian Terdahulu

Terdapat beberapa jurnal yang membahas analisis sentimen menggunakan algoritma yang akan digunakan oleh peneliti. Berikut Tabel 2.1 ini merupakan tabel penelitian terdahulu berisikan 8 jurnal internasional beserta hasil penelitiannya yang dianggap relevan oleh penulis dengan topik penelitian.

**Tabel 2.1** Penelitian Terdahulu

| No | Nama Jurnal | Judul | Tahun | Penulis | Hasil Penelitian |
|---|---|---|---|---|---|
| 1. | International Journal of Advanced Computer Science and Application (IJACSA) | Comparison of Naive Bayes and SVM Classification in Grid-Search *Hypermarater* Tuned and Non*Hypermarater* Tuned Healthcare Stock Market Sentiment Analysis [14]. | 2022, Vol. 13, no.12 pp.90-94, Scopus Q3. | KaiSiang Chong, Nathar Shah | Penelitian ini membandingkan kinerja klasifikasi Naïve Bayes dan SVM dalam analisis sentimen komentar saham perusahaan kesehatan di Bursa Malaysia dengan teknik Grid Search untuk penyetelan *hypermarater*. Hasil menunjukkan SVM memiliki akurasi lebih baik yaitu 85% dibandingkan dengan Naïve Bayes 68%. Penelitian ini menyatakan bahawa penyetelan *hypermarater* dapat meningkatkan kinerja model. |
| 2. | International Journal of Information Technology and Web Engineering (IJITWE) | Sentiment Analysis on Movie *Review*s *Dataset* Using Support Vector Machines and Ensemble Learning [15]. | Oktober 2022, Vol. 17, no.1 pp.1-23, Scopus Q3. | Razia Sulthana A., Jaithunbi A. K., Haritha Harikrishnan, Vijayakumar Varadarajan | Penelitian ini menerapkan teknik pemrosesan *machine learning* untuk menganalisis film menggunakan metode SVM, *dataset* terdiri dari 50 ribu ulasan IMDb dengan hasil akhir menunjukkan akurasi sebanyak 93.40% dengan menggunakan hiperparameter. |

| No | Nama Jurnal | Judul | Tahun | Penulis | Hasil Penelitian |
|---|---|---|---|---|---|
| | | | | | Kesimpulannya penelitian ini dapat dilakukan secara efektif dengan menggunakan *machine learning* metode SVM dan hiperparameter sehingga bisa diidentifikasikan kalimat kompleks dan sarkasme. |
| 3. | Applied Sciences (MDPI) | Application of Support Vector Machine (SVM) in the Sentiment Analysis of Twitter *Dataset* [16]. | Februari 2020, Vol 10, no.3 p. 11-25, Scopus Q2. | Kai-Xu-Han, Wei Chien, Chien-Ching Chiu, Yu-Ting Cheng | Penelitian ini mengusulkan fungsi FK-SVM berbasis Analisis Semantik Latent Probabilistik (PLSA) untuk meningkatkan hasil analisis sentimen menggunakan SVM. Metode FK-SVM ini dianggap lebih baik karena mencapai akurasi sebesar 87,20% jika dibandingkan dengan metode HIST-SVM (82,49%) dan PLSA-SVM. Kesimpulan yang didapat dengan menggunakan fungsi kernel Fisher lebih efektif dalam menangani informasi semantik laten dan dapat meningkatkan akurasi klasifikasi sentimen. |
| 4. | Journal of Engineering Science and Technology | Sentiment Analysis on Cosmetic Product in Sephora using Naïve Bayes Classifier [17]. | 2023, Vol 18, no.6 p. 11-21, Scopus Q3. | Fadly, Tri Basuki Kurniawan, Deshinta Arrova Dewi, Mohd Zaki Zakaria, Nazzatul Farahidayah Binti Mohd Nazziri | Penelitian ini menganalisis ulasan produk kosmetik di Sephora menggunakan algoritma Naïve Bayes dengan tujuan membantu pemantauan sentimen merk dan produk. Data dikumpulkan dari situs *web* Sephora dan diproses |

| No | Nama Jurnal | Judul | Tahun | Penulis | Hasil Penelitian |
|---|---|---|---|---|---|
| | | | | | menggunakan teknik pemrosesan bahasa alami, hasil algoritma menunjukkan akurasi 94,7% setelah pengujian dengan teknik validasi silang. Hasil produk yang diulas, merk Huda Beauty memiliki ulasan positif tertinggi, sedangkan Kat Von D menunjukkan ketidakseimbangan antara ulasan positif dan negatif. |
| 5. | Journal of Theoretical and Applied Information Technology | Sentiment Analysis for Tiktok Shop's Closure in Indonesia using Naïve Bayes Models and NLP [18]. | April 2024, Vol. 102, no.7 p. 2885-2894, Scopus Q4. | Henoch Juli Christanto, Steven Sondra Allen Widodo, Christine Dewi, Yerik Afrianto Singgalen, Dalianus Riantama, Andri Dayarana | Penelitian ini menggunakan teknik analisis sentimen dengan pemrosesan bahasa alami (NLP) untuk mengevaluasi reaksi publik terhadap penutupan Tiktok Shop di Indonesia. Data diambil dari Twitter dan dianalisis menggunakan algoritma Naïve Bayes. Hasil menunjukkan model Multinomial Naïve Bayes dengan labeling Textblob mencapai akurasi tertinggi sebesar 86%, yang mana artinya penutupan Tiktok Shop memiliki dampak signifikan terhadap industri e-commerce. |
| 6. | International Journal of Computer Applications | Sentiment Analysis of Tweets using SVM [19]. | November 2017, Vol. 177, no.5 p.25-29, Scopus Q3. | Munir Ahmad, Shahib Aftab, Iftikhar Ali | Penelitian ini menggunakan algoritma SVM dalam mengklasifikasikan tweet, *dataset* yang digunakan ada dua yaitu tentang mobil dan tentang produk |

| No | Nama Jurnal | Judul | Tahun | Penulis | Hasil Penelitian |
|---|---|---|---|---|---|
| | | | | | Apple. Kesimpulan dari penelitian ini adalah SVM memiliki performa bervariasi teergantung pada *dataset* yang digunakan, untuk *dataset* mobil nilai *Precision*, *Recall*, dan F-Measure adalah sebanyak 55.8%, 59,9%, dan 57,2%. Untuk *dataset* produk Apple memiliki nilai 70,2%, 71,2%, dan 69,9%. |
| 7. | Procedia Computer Science | Sentiment Analysis of Social Media Twitter with Case of Anti-LGBT Campaign in Indonesia Using Naive Bayes, Decision Tree, and Random Forest Algorithm [20]. | 2019, Vol. 161, no.1 p.756-772, Scopus. | Veny Amilia Fitri, Rachmadita Andreswari, Muhammad Azani Hasibuan | Penelitian ini menggunakan media sosial Twitter untuk melakukan analisis sentimen terhadap kampanye anti-LGBT dengan algoritma yang digunakan adalah Naïve Bayes, Decision Tree, dan Random Forest. Hasil yang diperoleh adalah akurasi dengan menggunakan algoritma Naïve Bayes lebih besar (86,43%) jika dibandingkan dengan algoritma lain yaitu Decision Tree dan Random Forest (82,91%), mayoritas berkomentar sentimen netral, dengan metode analisis meliputi pengumpulan data, *preprocessing*, klasifikasi, dan evaluasi. |
| 8. | ICIC Express Letters | Data Augmentation for Occlusion-Robust Traffic Sign | April 2024, Vol. 15, no.4 p.381- | Andrew Dineley, | Penelitian ini memiliki tujuan meningkatkan |

| No | Nama Jurnal | Judul | Tahun | Penulis | Hasil Penelitian |
|---|---|---|---|---|---|
| | | Recognition using Deep Learning [21]. | 388, Scopus Q4. | Friska Natalia, Sud Sudirman | akurasi pengenalan rambu lalu lintas yang terhalang oleh objek lain dengan menggunakan teknik augmentasi data dengan occlusion acak pada gambar rambu. Hasilnya terdapat peningkatan akurasi hingga 17% pada tingkat occlusion tinggi yaitu (61%-70%) menggunakan model GoogLeNet, meskipun ada penurunan kecil pada tingkat occlusion rendah. Augmentasi data signifikan meningkatkan kinerja model dalam kondisi occlusion parah. |

Melalui hasil penelitian terdahulu, telah didapatkan hasil dari penggunaan algoritma Naïve Bayes dan SVM pada analisis sentimental komentar di media sosial. Dari studi-studi tersebut, dapat disimpulkan bahwa algoritma SVM umumnya menunjukkan performa unggul dalam menangani data dengan pola kompleks, sedangkan Naïve Bayes lebih unggul pada data yang bersifat seimbang dan linier. Namun, kedua algoritma memiliki kelebihan dan kekurangan, SVM unggul dalam klasifikasi non-linier dan kasus data berukuran kecil hingga sedang, sementara Naïve Bayes cepat dan efisien untuk data besar namun sensitif terhadap korelasi antar fitur.

Penelitian oleh Ahmad [19] menganalisis sentimen pada data Twitter menggunakan algoritma Support Vector Machine (SVM). Dalam studi ini, penulis menunjukkan bahwa SVM mampu mengklasifikasikan tweet ke dalam tiga kategori sentimen (positif, netral, negatif) dengan akurasi mencapai 88,5%. Keunggulan SVM dalam menangani data teks pendek dan informal menjadi sorotan utama,

mengingat karakteristik tweet yang seringkali mengandung singkatan, simbol, dan struktur kalimat tidak baku.

Penelitian Veny [20] melakukan analisis sentimen terhadap kampanye Anti-LGBT di Twitter menggunakan algoritma Naïve Bayes, Decision Tree, dan Random Forest. Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes memberikan hasil paling akurat dengan nilai akurasi 86,43%, dibandingkan dengan dua algoritma lainnya. Penelitian ini menegaskan efektivitas Naïve Bayes dalam menangani data sosial media yang bersifat sangat opiniatif dan sarat emosi.

Penelitian oleh Chong & Shah [14] membandingkan kinerja algoritma Naïve Bayes dan SVM dalam menganalisis sentimen komentar saham perusahaan healthcare di Bursa Malaysia. Studi ini menekankan pentingnya *tuning hypermarater* melalui Grid Search, yang berhasil meningkatkan akurasi SVM hingga 85,65%, jauh di atas Naïve Bayes (68,70%). Penelitian ini menunjukkan bahwa performa algoritma sangat tergantung pada teknik optimasi dan konteks *dataset* yang digunakan.

Penelitian Sulthana [15]. menerapkan SVM dan *ensemble learning* untuk melakukan analisis sentimen terhadap ulasan film. Penelitian ini menggunakan *Chi-Square* untuk seleksi fitur dan *Grid Search* untuk optimasi parameter SVM. Hasilnya menunjukkan bahwa model SVM dengan bagging mampu memberikan performa lebih baik dibandingkan pendekatan standar. Penelitian ini mendukung bahwa penggunaan teknik *ensemble* dan pemilihan fitur yang tepat dapat meningkatkan akurasi model klasifikasi.

Penelitian Kai Xu [16]. mengembangkan metode baru dengan menggabungkan Probabilistic Latent Semantic Analysis (PLSA) dan kernel Fisher untuk meningkatkan akurasi SVM dalam analisis sentimen. Dengan pendekatan ini, dimensi semantik laten dari teks berhasil dimanfaatkan untuk memperkuat kinerja klasifikasi SVM. Studi ini memberikan kontribusi dalam mengatasi kelemahan SVM yang sering mengabaikan makna semantik dalam teks berbahasa alami.

Penelitian Christanto [18]. menggunakan varian Naïve Bayes (Multinomial, Bernoulli, dan Complement) serta TextBlob untuk menganalisis sentimen publik

terhadap penutupan TikTok Shop di Indonesia. Hasil tertinggi diperoleh dari kombinasi TextBlob dan CountVectorizer dengan akurasi mencapai 86,6%. Penelitian ini menyoroti pentingnya pemilihan fitur dan teknik *preprocessing* dalam menentukan keberhasilan klasifikasi sentimen dari media sosial.

Penelitian Andrew Dineley, Natalia, dan Sudirman [21] meneliti pengenalan rambu lalu lintas menggunakan deep learning dengan fokus pada data yang terhalang (*occluded*). Meskipun tidak berfokus pada teks, penelitian ini relevan karena menekankan pentingnya teknik data *augmentation*, *splitting* data dan pelatihan model pada kondisi data tidak ideal. Pendekatan ini dapat diadaptasi ke dalam pemrosesan teks sosial media yang memiliki noise, ketidakteraturan, dan konteks yang tidak eksplisit.

Tujuan penelitian ini berbeda dari studi sebelumnya, karena fokus utamanya adalah pada komentar Instagram terkait produk skincare D'Alba, yang kemudian dikaitkan dengan data penjualan di marketplace (Shopee dan Tokopedia). Mayoritas studi terdahulu hanya menilai performa algoritma berdasarkan akurasi atau F1-*score*, tanpa mengaitkan hasil sentimen terhadap indikator bisnis nyata seperti penjualan. Pada penelitian terdahulu umumnya menggunakan *dataset* dari media sosial Facebook atau Twitter, sedangkan pada penelitian ini menggunakan media sosial dari Instagram. Kebaruan penelitian ini adalah fokus kepada brand *skincare* yaitu D'Alba dengan Instagram dan *marketplace* sebagai sumber data dan kombinasi penggunaan algoritma SVM dan Naïve Bayes dalam analisis sentumen terkait *skincare*.

Urgensi dari penelitian ini sendiri adalah karena *skincare* merupakan sesuatu yang sangat sedang digemari, sehingga pentingnya opini dari masyarakat terkait produk-produk *skincare* yang marak terjual di Indonesia apakah bagus atau tidak dengan melihat ulasan dari salah satu media sosial. Hasil penelitian ini dapat memberikan wawasan bagi D'Alba untuk mengetahui produk yang lebih disukai berdasarkan hasil sentimen publik. Dari berbagai penelitian yang telah dibahas, dapat disimpulkan bahwa belum ada studi yang secara khusus menganalisis sentimen pengguna Instagram terhadap produk *skincare* tertentu dan

menghubungkannya dengan data penjualan *marketplace*. Dengan demikian, penelitian ini mengisi gap tersebut dengan pendekatan yang relevan dan terkini, serta memberikan kontribusi bagi brand dalam memahami konsumen secara lebih komprehensif.

## 2.2 Teori Penelitian

### *2.2.1 Skincare*

*Skincare* (perawatan kulit) merupakan suatu bahan yang mempunyai fungsi untuk menjaga kulit wajah agar tetap sehat dan melindungi dari berbagai masalah seperti jerawat, kulit kering, penuaan dini, dan lain-lain [22]. *Skincare* sendiri dapat berupa produk seperti serum, pelembap, pembersih, masker, dan masih banyak lagi. Berbagai jenis produk *skincare* beredar, namun perlu dipastikan apakah kandungan dalam produk tersebut aman untuk digunakan atau tidak [23].

Saat ini produk *skincare* sudah tersedia dengan berbagai bentuk dan formula sesuai dengan kebutuhan masing-masing jenis kulit dengan kandungan bahan alami seperti *white truffle*, *centella*, atau *niacinamide*. Melihat tingkatnya minat masyarakat dengan perawatan kulit, *skincare* tidak hanya menjadi suatu kebutuhan namun juga sudah menjadi gaya hidup untuk mengedepankan kesehatan. Oleh karena itu para pembeli biasanya akan melihat informasi mengenai *test*imoni dari produk tersebut melalui media sosial, khususnya media sosial *official* milik brand tersebut.

### 2.2.2 D'Alba

D'Alba merupakan salah satu merk *skincare* berasal dari Korea Selatan dibawah naungan perusahaan Bnoument.Co dengan CEO bernama Kim Hyun-Chul, perusahaan tersebut merupakan perusahaan kosmetik asal Korea Selatan yang berdiri sejak tahun 2010 [5]. D'Alba memiliki bahan utama yang berasal dari Italia karena bahan kandungannya menggunakan *white truffle*. *White truffle* dalam bahasa Italia adalah tartufo bianco yang merupakan jamur alami di hutan dan cukup sulit ditemukan sehingga harganya cukup mahal, bahan ini memiliki manfaat untuk meningkatkan

kelembapan kulit dan memberikan efek *anti-aging* dengan memudarkan garis halus di wajah kita [24].


**Gambar 2.1** White Truffle First Spray Serum

(Sumber: dalba.com) [5]

Gambar 2.1 merupakan salah satu produk dari D'Alba yang menggunakan *white truffle*. Melalui bahan alami ini produk D'Alba menjadi peluang bisnis yang bagus, di Indonesia produk ini dipasarkan secara resmi melalui akun Instagram @dalba_indonesia yang berisikan informasi mengenai produk dan ulasan konsumen. Dengan kelebihan bahan utama, D'Alba telah meningkatkan reputasinya sebagai salah satu *skincare* premium yang diminati di pasar internasional.

### 2.2.3 Instagram

Instagram adalah salah satu platform media sosial yang banyak digunakan untuk membagikan gambar, video, dan pendapat. Instagram merupakan tempat dimana para pengguna sering membagikan ulasan terkait suatu hal yang digunakan [8]. Oleh sebab itu, media sosial ini dapat memberikan data berbentuk komentar berupa ulasan dari para pengguna terkait produk yang digunakan sehingga bisa dipahami persepsi konsumen.

Instagram diluncurkan pada tahun 2010 oleh Kevin Systrom dan Mike Krieger, pada awalnya aplikasi ini hanya digunakan untuk berbagi foto, namun seiring berjalannya waktu Instagram dapat dijadikan alat pemasaran yang berfokus pada visual seperti makanan, perawatan kulit, fashion, dan masih banyak lainnya [25]. Instagram memberikan *insight* mengenai opini

dan persepsi masyarakat terkait suatu hal sehingga bisa digunakan untuk membangun citra suatu hal.

### 2.2.4 *Marketplace* (Shopee dan Tokopedia)

*Marketplace* merupakan suatu platform yang mempertemukan penjual dan pembeli sehingga dapat melakukan transaksi jual beli produk dan jasa secara efisien dengan cara para pengguna bisa membandingkan atau mencari berbagai produk dari berbagai penjual tanpa harus datang ke toko tersebut [26]. *Marketplace* memiliki keunggulan tersendiri dikarenakan memungkinkan untuk banyaknya penjual yang menawarkan produk sama namun dengan harga berbeda. Berdasarkan latar belakang, Shopee dan Tokopedia sendiri merupakan dua *marketplace* terbesar di Indonesia yang memiliki pengaruh signifikan.

Shopee mulai hadir pada tahun 2015 oleh Sea Group, kemudian berkembang pesat menjadi salah satu platform belanja daring di Asia Tenggara yang termasuk di dalamnya ada Indonesia dengan berbagai fitur dan promo seperti gratis ongkir dan lainnya, hal tersebut sesuai karena Shopee mengusung pendekatan dengan *mobile-first* [27].

Sementara itu, Tokopedia didirikan lebih dulu pada tahun 2009 oleh William Tanuwijaya dan Leontinus Alpha Edison, Tokopedia mengusung visi meratakan ekonomi digital sehingga lebih berfokus pada budidaya UMKM dengan menyediakan berbagai kategori produk mulai dari kebutuhan rumah tangga hingga produk kecantikan [28].

### 2.2.5 Analisis Sentimen

Analisis sentimen merupakan sebuah proses mendapatkan berbagai sumber data dari internet dan platform media sosial. Teknik yang digunakan berupa mengekstrak dan menganalisis opini dan perasaan yang terkandung dalam teks, seperti ulasan dari komentar online [29]. Tujuan analisis sentimental adalah untuk memahami apakah suatu data tersebut memiliki sentimen yang positif, negatif, atau netral, dengan begitu nantinya pembaca atau para pemangku kepentingan dapat memperoleh inti atau masukan dari

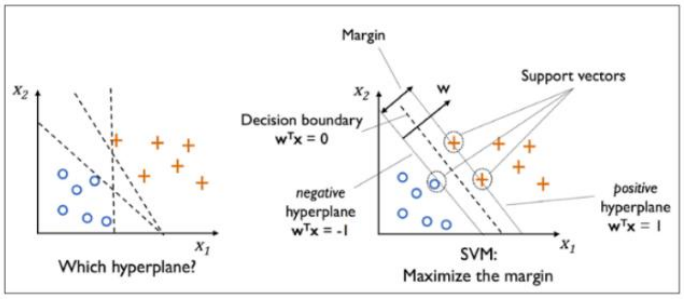
ulasan-ulasan masyarakat yang sudah diberikan secara efisien. Berikut dibawah ini merupakan beberapa langkah yang dapat dilakukan ketika melakukan analisis sentimen yaitu [30]:

1. **Pengumpulan Data:**
   Data yang mengandung opini atau sentimen yang telah dikumpulkan dari berbagai sumber seperti media sosial dan survei.
2. **Proses Teks:**
   Teks diproses untuk dibersihkan dari tanda baca atau kata-kata yang tidak relevan, teknik yang biasa digunakan dengan bantuan *machine learning*.
3. **Ekstrasi Fitur:**
   Melihat fitur-fitur apa yang relevan dari teks atau opini yang telah dikumpulkan agar dapat sesuai dengan topik.
4. **Klasifikasi Algoritma:**
   Algoritma untuk mengklasifikasikan teks kedalam kategori sentimen yaitu positif, negatif, atau netral, pada penelitian ini digunakan algoritma SVM dan Naïve Bayes.
5. **Evaluasi Model:**
   Model yang sudah dilakukan klasifikasi diuji untuk mengetahui seberapa akurat algoritma tersebut mengklasifikasikan sentimen.
6. **Analisis Hasil:**
   Hasil dari analisis sentimen yang dilakukan dipahami untuk melihat bagaimana persepsi masyarakat sehingga bisa diambil tindakan berdasarkan dari wawasan yang sudah diperoleh.

## 2.3 *Framework* dan Algoritma Penelitian

### 2.3.1 Support Vector Machine (SVM)

SVM merupakan algoritma *machine learning* yang digunakan untuk klasifikasi dan regresi. SVM akan mencari hyperplane optimal untuk memisahkan data ke dalam kelas berbeda dengan margin maksimum [31]. Berikut Gambar 2.5 ini merupakan gambaran dasar cara kerja algoritma SVM.




**Gambar 2.2** Hyperplane Negatif dan Positif

(Sumber: medium.com) [33]

Pada gambar 2.5 terdapat dua bagian yang akan menjelaskan bagaimana SVM menentukan hyperplane terbaik untuk memisahkan dua kelas data (lingkaran biru dan tanda plus oranye). Untuk bagian kiri menunjukkan beberapa hyperplane yang dapat digunakan untuk memisahkan dua kelas, sedangkan untuk bagian kanan adalah proses dimana SVM memilih hyperplane dengan memaksimalkan margin antara kedua kelas tersebut. Margin merupakan jarak antara hyperplane dan dat terdekat dari setiap kelas (*support vectors*), hyperplane terbaik adalah yang memberikan margin terbesar.

Hyperplane sendiri merupakan batas keputusan untuk memisahkan data dari berbagai kelas dalam ruang dimensi tinggi. Titik data yang paling dekat dengan hyperplane disebut *support vectors* dan berfungsi untuk menentukan posisi dari hyperplane tersebut [16]. Tujuan utama SVM dalam menemukan hyperplane tidak hanya untuk memisahkan kelas namun juga untuk memaksimalkan jarak antara kelas tersebut sehingga kemampuannya dapat meningkat ketika generalisasi model yang sebelumnya tidak pernah dilihat.

SVM memiliki keunggulan dalam bekerja dengan baik pada data yang berdimensi tinggi dimana efektivitasnya pada jumlah fitur lebih besar daripada jumlah sampel. Namun, SVM bekerja kurang efisien ketika terdapat banyak *noise* (komponen data yang salah, tidak relevan, dan menyimpang dari pola yang sebenarnya) di dalam data [32]. Algoritma ini digunakan untuk mengklasifikasikan ulasan Instagram pengguna D'Alba sebagai sentimen positif, negatif, atau netral.

### 2.3.2 Naïve Bayes

Naive Bayes atau merupakan algoritma yang biasa digunakan dalam *machine learning* dan statistik. Algoritma ini didasarkan dari teori Bayes yang digunakan untuk menghitung probabilitas kelas secara sesuai berdasarkan probabilitas yang sudah diketahui sebelumnya. Algoritma ini disebut Naïve dikarenakan mengasumsikan independensi antara fitur yang digunakan dalam pemodelan dan memiliki asumsi bahwa setiap fitur tidak terkait dengan fitur lainnya [33].

$$\frac{\boldsymbol{P(Ck \mid x) = P(x)P(Ck) \cdot \prod i = 1nP(xi \mid Ck)x}}{\boldsymbol{P(x)}}$$

Gambar 2.6 menjelaskan probabilitas bersyarat yang digunakan untuk menghitung kemungkinan terjadinya kejadian A namun dengan syarat bahwa kejadian B sudah terjadi. Rumus ini merupakan dasar dari Teorema Bayes untuk pengklasifikasian data dalam algoritma Naïve Bayes. Berdasarkan rumus akan dihubungkan dua kejadian yang mungkin memiliki keterkaitan dengan memperhitungkan setiap probabilitas kejadian masing-masing [34].

Salah satu asumsi utama dari Naïve Bayes adalah fitur yang digunakan dalam model klasifikasi adalah independen satu dengan yang lain, meskipun hal tersebut hanya asumsi namun dalam banyak kasus Naïve Bayes sudah menghasilkan hasil yang baik. Naïve Bayes dapat diterapkan dalam berbagai jenis klasifikasi termasuk klasifikasi teks, gambar, dan yang lainnya. Kelas yang memiliki probabilitas tertinggi maka dianggap sebagai prediksi kelas yang benar, untuk mengukur kinerjanya dapat digunakan metrik seperti F1-*score*.

Naïve Bayes sendiri cocok digunakan dalam proses analisis sentimen karena cara klasifikasi datanya yang cepat dan sederhana sehingga cocok untuk *dataset* yang besar, namun kekurangannya adalah performa akan menurun jika asumsi dari independensi tidak terpenuhi [33]. Algoritma ini digunakan untuk perbandingan performa dengan algoritma SVM dalam menganalisis sentimen ulasan *skincare* D'Alba di Instagram.

### 2.3.3 CRISP-DM


**Gambar 2.3** CRISP-DM Diagram Proses

(Sumber: mmsi.binus.ac.id) [37]

CRISP-DM atau *cross-industry standard process for data mining* merupakan salah satu proses datamining yang digunakan secara luas di kalangan industri karena keunggulannya dalam menyelesaikan berbagai persoalan [35], dengan menggunakan CRISP-DM penelitian menjadi lebih terstruktur dan memiliki kerangka kerja. Terdapat beberapa tahap dalam model CRISP-DM yaitu:

1. **Business Understanding:**

   Merupakan proses memahami tujuan bisnis dan kebutuhan dari perusahaan, dengan mengidentifikasi tujuan bisnis untuk mengetahui masalah apa yang akan dipecahkan dan aspek lain yang relevan.

2. **Data Understanding:**

   Merupakan proses mengumpulkan data awal untuk memahami karakteristiknya dengan cara eksplorasi awal data tersebut kemudian di identifikasi masalah pada data sehingga didapat pemahaman awal terhadap hubungan antara variabel.

3. **Data Preparation:**

   Merupakan proses mempersiapkan data untuk analisis dengan melakukan pembersihan data, menggabungkan berbagai sumber data dan mengonversi data tersebut.

4. **Modeling:**

   Merupakan proses membangun model atau metode analisis data denga melakukan pemilihan dan penerapan teknik pemodelan dan membagi data untuk pelatihan dan pengujian model sehingga nantinya dapat dilakukan penyesuaian terhadap parameter model.

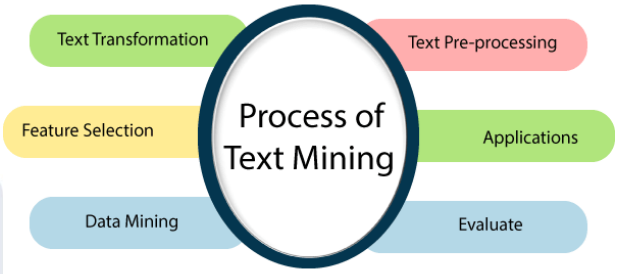
5. **Evaluation:**

   Merupakan proses mengevaluasi kinerja model terhadap tujuan bisnis dengan menguji kinerja dan validasi hasil. Model dievaluasi dengan membandingkan metrik seperti akurasi dan *recall*.

6. ***Deployment*:**

   Merupakan proses mengimplementasikan hasil analisis ke lingkungan produksi dengan cara membuat laporan dan integrasis model ke sistem produksi sehingga dapat dilaksanakan strategi implementasi.

### *2.3.4 Text-Mining*




**Gambar 2.4** Proses *Text mining*

(Sumber: javatpoint) [38]

*Text mining* merupakan proses mengekstrak informasi dari data teks yang tidak terstruktur. Teknik ini dapat digunakan untuk menganalisis berbagai teks, seperti dalam laporan, berita, artikel, komentar media sosial, dan masih banyak lagi sehingga nantinya dapat ditemukan wawasan dari penggalian informasi tersebut. *Text mining* digunakan karena dapat membantu memecahkan teks yang kompleks untuk diidentifikasikan tren atau polanya. Hal ini meliputi beberapa tahap seperti berikut [36]:

1. ***Text Pre-processing*:**

   Data dikumpulkan dari berbagai sumber, namun pada penelitian ini menggunakan sumber data yang berasal dari komentar (ulasan) para pengguna di Instagram terkait produk D'Alba. Kemudian dilakukan langkah-langkah untuk mempersiapkan teks seperti menghapus tanda baca, mengubah kata menjadi bentuk dasarnya seperti "memiliki" menjadi "milik" (*stemming*), memecah teks menjadi sebuah kata-kata (tokenisasi), mengubah semua teks menjadi huruf kecil (case folding), menghapus tanda baca, dan menghilangkan kata-kata yang tidak memiliki arti contoh "dan" "atau" (stop words).

2. ***Text Transformation*:**

   Data teks yang sudah di bersihkan diubah menjadi representasi numerik dengan metode vektorisasi, dalam penelitian ini digunakan metode TF-IDF untuk memberikan bobot berdasarkan frekuensi kemunculan kata.

3. ***Feature Selection*:**

   Dari hasil vektorisasi, kata yang paling relevan akan dipilih untuk di analisis lebih lanjut. Teknik ini bertujuan untuk mengurangi dimensi dari data dan meningkatkan efisiensi algoritma.

4. ***Data mining*:**

   Tahap ini akan menerapkan kedua algoritma yaitu SVM dan Naïve Bayes untuk membangun model klasifikasi, lalu kedua algoritma dibandingkan dari segi akurasi, presisi, dan metrik evaluasi untuk menentukan algoritma mana yang lebih efektif dalam menganalisis sentimen.

5. ***Evaluation*:**

   Setelah model dibangun, hasil klasifikasi di evaluasi menggunakan metrik seperti akurasi, *precision*, *recall*, dan F-1 *score* untuk memastikan bahwa model memiliki kinerja yang baik dalam mengidentifikasi sentimen komentar konsumen.

6. ***Application*:**

   Hasil analisis kemudian dapat diterapkan untuk mengidentifikasi persepsi konsumen terhadap produk D'Alba apakah cenderung positif, negatif, atau netral. Informasi ini dapat digunakan untuk meningkatkan strategi pemasaran dan memahami kebutuhan konsumen.

Teknik ini dilakukan agar data teks ulasan dari Instagram bisa diambil informasi berharga dari teks yang tidak terstruktur sehingga bisa lanjut dianalisis menggunakan algoritma SVM dan Naïve Bayes. Penerapan *text mining* dapat mengevaluasi persepsi konsumen secara objektif berdasarkan data sehingga hasil penelitian dapat memberikan manfaat bagi D'Alba dan konsumen.

### 2.3.5 Vektorisasi

Vektorisasi merupakan proses ketika mengubah teks menjadi bentuk numerik (vektor) sehingga bisa dipahami oleh algoritma *machine learning*. Dalam konteks analisis sentimen, vektorisasi bertujuan untuk memahami hubungan antar kata dalam teks dengan cara mengonversi kalimat menjadi representasi numerik [14]. Proses ini penting dilakukan karena penggunaan algoritma SVM dan Naïve Bayes menggunakan data numerik bukan data mentah, berikut dibawah ini merupakan beberapa metode umum dari vektorisasi:

1. ***Bag of Words*:**

   Bag Of Words atau BoW merupakan metode dengan cara kerja mengumpulkan kata unik berdasarkan frekuensi dari sumber data tanpa memperhatikan urutan kata tersebut. Misal, jika ada 100 kata unik dalam data, maka akan langsung direpresentasikan sebagai vektor dengan panjang 100.

2. ***Term Frequency-Inverse Document Frequency* (TF-IDF):**

   TF-IDF merupakan metode dengan cara kerja memberikan bobot pada tiap kata berdasarkan frekuensi kemunculannya di data.

Metode ini lebih efektif dibandingkan BoW untuk menekan kata-kata spesifik namun tidak terlalu sering muncul.

$$TF - IDF = TF \, x \, IDF$$

(Sumber: edinesia.com) [37]

3. ***Word Embedding*:**

   Metode ini menggunakan representasi kata dalam bentuk vektor dimensi rendah seperti Fast Text, umumnya menangkap hubungan semantik antar kata seperti sinonim atau konotasi.

Pada penelitian ini akan dilakukan proses vektorisasi dengan metode TF-IDF untuk mengubah komentar dari Instagram menjadi numerik sehingga kedua algoritma dapat memprosesnya. Metode ini dipakai karena keunggulannya dalam memberikan bobot yang lebih tinggi pada kata-kata penting dan lebih berfokus pada kata-kata bermakna [37].

Misalnya dalam ulasan terdapat kata "bagus" dan "*glowing*" akan memiliki frekuensi tinggi jika kata tersebut sering muncul namun kata-kata seperti "dan" atau "itu" akan memiliki bobot lebih kecil karena kurang informatif, kesimpulannya dengan metode TF-IDF dapat memprioritaskan kata yang relevan dan mengurangi dampak kata-kata umum (stop words).

### *2.3.6 Confusion Matrix*

Confusion Matrix merupakan sebuah alat evaluasi yang digunakan di *machine learning*, khususnya pada masalah klasifikasi memungkinkan dalam memahami kinerja model dengan membandingkan prediksi model dengan nilai aktual. Matriks sendiri terdiri dari empat komponen utama [38]:

1. ***True Positive* (TP):**

   Prediski positif benar, model memprediksi kelas positif dan hal itu memang benar adanya.

2. ***True Negative*** **(TN):**

   Prediksi negatif yang benar, model memprediksi kelas negatif dan hal itu memang benar adanya.

3. ***False Positive*** **(FP):**

   Prediksi positif yang salah, model memprediksi kelas positif namun seharusnya negatif.

4. ***False Negative*** **(FN):**

   Prediksi negatif yang salah, model memprediksi kelas negatif padahal seharusnya positif.

Dengan penggunaan confusion matrix, penghitungan metrik seperti akurasi, presisi, *recall*, dan F1-*score* dapat memberikan wawasan lebih mendalam mengenai kinerja model dibandingkan pengandalan akurasi. Menerapkan confusion matrix dengan benar sangat penting dalam proses mengevaluasi kinerja model klasifikasi.

## 2.4 *Tools* dan *Software* Penelitian

### 2.4.1 Microsoft Excel

Microsoft Excel merupakan perangkat lunak spreadsheet yang dikembangkan oleh Microsoft. Excel merupakan program yang fungsi utamanya untuk mengolah data yang berupa angka menggunakan spreadsheet dalam penyajian baris serta kolom untuk mengeksekusi perintah [39]. Alasan peneliti juga menggunakan Excel karena aplikasi tersebut mudah mengimpor dan menggabungkan data dengan dokumen lain, kemudian Excel juga mendukung manipulasi data yang kuat seperti sortir data, menyalin data, menggabungkan data dari berbagai sumber, dan sebagainya yang akan memudahkan peneliti dalam menganalisis data tersebut.

### 2.4.2 *Anaconda*

*Anaconda* merupakan sebuah perangkat lunak open source yang fokus terhadap pemrograman dan analisis data. Hal ini mencakup berbagai paket dan alat yang berguna untuk menganalisis data dalam bahasa pemrograman

*Python*. *Anaconda* sendiri menyediakan paket bernama Conda yang memungkinkan para pengguna dapat menginstal, mengelola, dan menghapus paket-paket tersebut dengan mudah, hal tersebut juga tentu berguna dalam mengelola berbagai dependensi software [40]. *Anaconda* juga menyediakan berbagai paket seperti NumPy, pandas, Matplotlib, dan berbagai paket lainnya yang menjadikan *Anaconda* sebagai salah satu platform yang kuat dalam menganalisis data dan pemodelan statistik.

### 2.4.3 *Jupyter notebook*

Salah satu fitur yang cukup sering digunakan di *Anaconda* yaitu *Jupyter notebook*. *Jupyter notebook* merupakan tempat dimana pengguna dapat menulis kode dalam bahasa *Python* secara berurutan dalam sel dan menjalankannya sehingga akan didapat hasil yang interaktif untuk eksplorasi dan pemodelan data [41]. Sebagai pengguna juga dapat menyimpan format Notebook dalam format yang bisa dibagikan ke semua orang, sehingga dapat digunakan juga untuk analisis laporan atau catatan ilmiah.

### 2.4.4 *Python*

*Python* merupakan sebuah bahasa pemrograman tingkat tinggi yang sering digunakan untuk melakukan analisis data, membuat aplikasi, dan lain sebagainya [42]. *Python* memiliki bahasa yang mudah dipahami dan mirip dengan bahasa manusia, sehingga hal tersebut memudahkan peneliti dalam menulis dan membaca kode *Python*. Selain itu alasan dari peneliti memilih bahasa *Python* untuk digunakan karena *Python* memiliki berbagai framework yanng luas serta beragam, adapun penggunaannya yang sangat fleksibe dan dapat digunakan dalam berbagai platform seperti Windows, macOS, dan Linux.