BAB III
METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Penelitian ini berfokus pada analisis sentimen pengguna Instagram terhadap produk skincare D'Alba, serta hubungannya dengan data penjualan di marketplace Shopee dan Tokopedia. D'Alba merupakan merek skincare asal Korea Selatan yang dikenal luas dengan inovasi penggunaan bahan premium white truffle dari Italia. Produk ini diklaim memiliki berbagai manfaat untuk kesehatan kulit seperti anti-aging, pencerahan, dan hidrasi intensif. Seiring meningkatnya minat konsumen Indonesia terhadap skincare Korea, D'Alba menjadi salah satu brand yang cukup populer di kalangan pengguna media sosial, terutama di Instagram.

Instagram dipilih sebagai platform utama dalam penelitian ini karena merupakan salah satu media sosial paling aktif di Indonesia dengan tingkat interaksi tinggi, khususnya dalam bentuk komentar pengguna terhadap produk.Penelitian ini menggunakan dua algoritma klasifikasi sentimen, yaitu Support Vector Machine (SVM) dan Naïve Bayes, yang masing-masing akan dievaluasi untuk mengetahui performa akurasinya. Untuk pemodelan data dan tahapan analisis, digunakan framework CRISP-DM (*Cross Industry Standard Process for Data mining*) karena struktur alurnya yang sistematis dan fleksibel.

Selain itu, penelitian ini juga menelaah apakah sentimen yang muncul di media sosial memiliki korelasi terhadap penjualan produk D'Alba di marketplace Shopee dan Tokopedia. Kedua platform ini merupakan e-commerce terbesar di Indonesia berdasarkan jumlah pengunjung dan transaksi. Data yang dianalisis mencakup jumlah produk terjual, rating, dan harga produk.

Jika ditemukan bahwa sentimen positif memiliki hubungan kuat dengan tingginya penjualan dan rating, maka dapat disimpulkan bahwa opini di media sosial memiliki pengaruh signifikan terhadap keputusan pembelian. Sebaliknya, jika penjualan tetap tinggi meski sentimen dominan netral atau negatif, maka dapat diindikasikan bahwa faktor lain seperti promosi, loyalitas brand, atau strategi marketplace lebih berpengaruh. Dengan memahami gambaran umum ini,

diharapkan penelitian dapat memberikan wawasan yang relevan bagi industri kosmetik dan e-commerce, khususnya dalam memanfaatkan media sosial sebagai indikator performa pasar dan strategi pemasaran.

3.2 Metode Penelitian

Pendekatan penelitian yang digunakan adalah pendekatan kuantitatif, karena data yang dianalisis berbentuk numerik dan memerlukan proses perhitungan statistik serta pengujian performa model menggunakan metrik evaluasi. Metode yang digunakan pada penelitian ini adalah CRISP-DM (*cross-industry standard process for data mining*) sebagai acuan dalam melakukan analisis sentimen *skincare* D'Alba. Framework ini dipilih karena memiliki alur yang sistematis dan terstruktur dalam proses *Data mining*. Berikut dibawah ini merupakan tahapan penelitian:

1. ***Business Understanding*:**

   Pada penelitian ini tujuan dilakukan analisis sentimen komentar adalah agar bisa memberikan wawasan terkait produk D'Alba pada perusahaan maupun pada konsumen serta melihat kaitannya dengan penjualan di marketplace dan menentukan algoritma mana yang lebih efektif dalam mengklasifikasi sentimen. Informasi ini diharapkan dapat membantu mendukung strategi pemasaran dan pengembangan produk untuk D'Alba.

2. ***Data Understanding*:**

   Data ulasan dari Instagram dikumpulkan dengan metode *scraping* menggunakan pihak ketiga yaitu *website exportcomments*, untuk dianalisis dan dipahami karakteristiknya seperti teks pendek, bahasa informal, dan penggunaan emoji, jumlah data, dan mengidentifikasi noise. Data komentar Instagram diambil dalam periode waktu Januari 2021 hingga Januari 2025 menggunakan kata kunci tagar #dAlbaIndonesia. Sedangkan untuk pengambilan data Tokopedia dan Shopee digunakan pihak ketiga yaitu *API Shopee Chat Tuesday* dan Tokopedia *Data Scraper*.

3. ***Data Preparation*:**

   Proses data preparation dilakukan dengan teknik *text mining* seperti tokenisasi, *stemming*, cleaning, *stopword removal*, dan labeling.

4. ***Modeling*:**

   Algoritma SVM dan Naïve Bayes diterapkan untuk memodelkan data ulasan dan mengklasifikasikannya menjadi sentimen positif, negatif, atau netral. Selain itu data akan dikonversi menjadi vektor numerik menggunakan teknik vektorisasi TF-IDF (Term Frequency-Inverse Document Frequency), kemudian dilakukan teknik *train-test* dengan 80% untuk data *train*, dan 20% untuk data *test*. Lalu metrik seperti akurasi, *recall*, *precision*, dan F-1 digunakan untuk membandingkan kedua performa algoritma.

5. ***Evaluation*:**

   Evaluasi dilakukan untuk menentukan apakah model sudah sesuai tujuan atau belum dengan cara membandingkan kedua akurasi algoritma dan menganalisis kekuatan serta kelemahan tiap algoritma berdasarkan hasil klasifikasi yang dilakukan dan melihat hasil sentimennya.

6. ***Deployment*:**

   Hasil penelitian disajikan dalam bentuk laporan akademik dan *website streamlite* yang mencakup wawasan mengenai persepsi konsumen terhadap D'Alba sehingga bisa dijadikan rekomendasi untuk pengembangan produk serta wawasan mengenai keefektifan algoritma SVM dan Naïve Bayes dalam menganalisis sentimen.

3.2.1 Perbandingan Metode

Metode pengolahan data yang dilakukan pada penelitian ini adalah CRISP-DM sebagai kerangka utama. Untuk memberikan perbandingan lebih lanjut, berikut Tabel 3.1 ini merupakan perbandingan dari kedua framework yang sering digunakan dalam *data mining* yaitu CRISP-DM [43] dan KDD (*Knowledge Discovery in Databases*) [44]:

Tabel 3.1 Perbandingan Teknik *Data mining*

Indikator	CRISP-DM	KDD
Fokus	Menyelesaikan masalah berbasis bisnis.	Menemukan pengetahuan baru dari hasil data.
Konteks	Mempertimbangkan tujuan bisnis.	Kurang menonjolkan aspek bisnis.
Penerapan	Lebih cocok karena lebih fleksibel dan relevan	Lebih cocok dalam konteks evaluasi yang tidak spesifik.

Indikator	CRISP-DM	KDD
	berdasarkan tujuan penelitian.	
Struktur	Rinci dan terarah dengan 6 tahap.	Lebih fokus pada eksplorasi dan evaluasi data.

Kesimpulan berdasarkan Tabel 3.1 perbandingan metode, alasan peneliti lebih memilih menggunakan framework CRISP-DM karena dapat membantu penelitian berjalan secara rinci dan terarah, semua tahapan jelas dan sesuai dengan tujuan penelitian, serta tidak hanya berfokus pada eksplorasi data namun juga pada evaluasi model dan hasil analisis sentimen dalam konteks bisnis.

3.2.2 Perbandingan Algoritma

Penelitian ini menggunakan dua algoritma yaitu SVM dan Naïve Bayes dalam proses perbandingan performanya melakukan analisis sentimen terhadap komentar Instagram para konsumen tentang *skincare* D'Alba. Berikut Tabel 3.2 ini merupakan perbandingan antara kedua algoritma tersebut berdasarkan karakteristiknya [14].

Tabel 3.2 Perbandingan Algoritma

Indikator	SVM	Naïve Bayes
Akurasi	Lebih tinggi, terutama untuk data kompleks.	Memadai untuk data yang sederhana.
Kecepatan Komputasi	Lebih lambat jika digunakan pada *dataset* besar.	Lebih cepat dan efisien.
Penerapan	Bagus untuk menangani sentimen yang kompleks.	Kurang bagus pada data yang kompleks.
Implementasi	Lebih kompleks.	Cepat dan sederhana.

Kesimpulannya algoritma SVM lebih unggul dalam hal akurasi dan kemampuannya untuk menangani data kompleks, sehingga cocok untuk klasifikasi sentimen yang perlu banyak pola. Algoritma Naïve Bayes lebih cocok digunakan untuk kasus yang sederhana dengan jumlah data besar. Peneliti akan membandingkan tentang efektivitas kedua algoritma tersebut dalam proses analisis sentimen ulasan Instagram di produk D'Alba untuk memberikan rekomendasi pada pembaca.

3.3 Teknik Pengumpulan Data

Penelitian ini menggunakan data primer yang dikumpulkan melalui metode *scraping* data online, yaitu pengambilan data otomatis dari platform digital. Data komentar Instagram diperoleh menggunakan tools pihak ketiga bernama ExportComments, yang mengambil komentar dari akun resmi @dalba_indonesia serta dari akun beauty influencer yang membahas produk D'Alba. Untuk data marketplace, digunakan API Chat Tuesday (Shopee) dan Tokopedia Data Scraper. Berikut contohnya:

Tabel 3.3 Contoh Hasil *Scraping*

username	profile_id	date	comment	source
beautylover23	35293168712	09/08/24 05:08:16	Produk ini bikin kulitku glowing banget! 💖	Link IG
daldal_lovers	5716051177	09/08/24 05:17:08	Kok aku malah jadi breakout ya setelah pakai?	Link IG
skincare.addict	32057967	09/08/24 06:19:57	Penasaran bgt sama spray serumnyaa 😘	Link IG
nama_produk	**harga_produk**	**rating**	**quantity_sold**	**-**
d Alba White Truffle First Spray Serum	Rp. 225.000	5	3 terjual	-

3.3.1 Periode Pengambilan Data

Data dikumpulkan dari akun Instagram mengenai skincare D'Alba dalam rentang waktu selama 3 tahun yaitu pada tanggal 1 Januari 2021 sampai dengan 1 Desember 2023, namun untuk data uji diambil dalam rentang waktu selama 1 tahun yaitu pada tanggal 1 Januari 2024 sampai dengan 1 Januari 2025 dengan kurang lebih telah ada 234 unggahan (data latih) dan 80 unggahan (data uji) mengenai produk D'Alba yang tiap postingan berisi komentar dari para konsumen. Kemudian untuk data dari *marketplace* telah terkumpul 798 data dalam rentang waktu ketika produk D'Alba masih aktif dijual.

3.4 Teknik Analisis Data

Analisis data dalam penelitian ini dilakukan melalui beberapa tahap seperti dibawah ini.

1. **Pengumpulan Data**

   Data yang digunakan merupakan komentar di Instagram yang membahas produk D'Alba, kemudian data tersebut dikumpulkan yang hasilnya berupa file Excel dengan periode pengambilan waktu selama 3 tahun yaitu 1 Januari 2021 sampai dengan 1 Desember 2023 (data latih) sedangkan data uji periode 1 tahun yaitu 1 Januari 2024 sampai dengan 1 Januari 2025. Untuk data *marketplace*, *scraping* data diambil tanpa rentang waktu selama produk tersebut masih terjual.

2. **Pra-proses Data**

   Data komentar dibersihkan dan diproses menggunakan teknik *preprocessing*.

3. **Pemodelan (*Modeling*)**

   Setelah data siap dianalisis, dilakukan proses *splitting* data dan vektorisasi TF-IDF maka akan dilakukan pemodelan dengan dua algoritma yaitu SVM dan Naïve Bayes.

4. **Evaluasi Model (*Evaluation*)**

   Kedua algoritma tersebut dibandingkan dengan acuan *Confusion Matrix* yang mencakup metrik akurasi (seberapa sering model mengklasifikasikan data dengan benar), *precision* (ketepatan model dalam mengklasifikasikan sentimen), *recall* (kemampuan model dalam menangkap seluruh komentar), F-1 *Score* (rata-rata dari *precision* dan *recall* untuk menyeimbangkan kinerja model).

5. **Interpretasi Hasil**

   Peneliti akan membandingkan algoritma mana yang lebih baik dalam menganalisis sentimen *skincare* D'Alba berdasarkan hasil metrik. Ketika proses interpretasi hasil dilakukan, terdapat alat analisis data yang akan digunakan, berikut perbandingan kedua alat yang digunakan untuk *data mining* [45]:

Tabel 3.4 Perbandingan Tools Analisis Data

Indikator	*Jupyter notebook*	R-Studio
Analisis Statistik	Memiliki berbagai library yang kuat untuuk di lakukan analisis data.	Memiliki alat statistik yang lebih lengkap.

Indikator	*Jupyter notebook*	R-Studio
Fokus	Lebih fokus terhadap berbagai ilmu data seperti *machine learning* dan pengembangan perangkat.	Lebih fokus pada analisis statistik.
Kelebihan	Mudah digunakan karena memiliki sintaks yang lebih sederhana	Mudah digunakan karena lebih fokus pada statistic dan analisis data.
Kekurangan	Terlalu banyak sintaks yang beragam sehingga sulit untuk digunakan bagi pengguna baru.	Hanya fokus pada analisis statistik dan visualisasi data.

Berdasarkan hasil perbandingan, peneliti memilih untuk menggunakan *Jupyter notebook* (*Python*) dikarenakan memiliki berbagai library yang sangat mendukung dalam melakukan analisis data, selain itu, *Jupyter notebook* juga lebih mudah dipelajari dan dipahami karena sering digunakan oleh peneliti semasa kuliah.

3.4.1 Rumusan Hipotesis

Penelitian ini didasarkan pada beberapa dugaan awal yang dirumuskan dalam bentuk hipotesis. Hipotesis ini akan diuji melalui proses klasifikasi sentimen menggunakan kedua algoritma tersebut serta analisis korelasi antara hasil sentimen dengan data penjualan. Berikut merupakan hipotesis dari penelitian ini:

Tabel 3.5 Hipotesis Penelitian

Rumusan Masalah	Hipotesis Nol (H_0)	Hipotesis Alternatif (H_1)
Bagaimana sentimen pengguna Instagram terhadap produk skincare D'Alba?	Tidak terdapat sentimen dominan dalam komentar pengguna Instagram terhadap produk D'Alba.	Terdapat sentimen dominan (positif, negatif, atau netral) dalam komentar pengguna Instagram terhadap produk D'Alba.
Bagaimana perbandingan performa algoritma SVM dan Naïve Bayes dalam mengklasifikasikan sentimen?	Tidak terdapat perbedaan performa klasifikasi antara algoritma SVM dan Naïve Bayes.	Algoritma SVM memiliki performa klasifikasi yang lebih baik dibandingkan algoritma Naïve Bayes.
Apakah terdapat korelasi antara sentimen komentar pengguna Instagram terhadap produk D'Alba dengan jumlah penjualan di Shopee dan Tokopedia?	Tidak terdapat korelasi antara sentimen komentar di Instagram dengan jumlah penjualan produk D'Alba di marketplace.	Terdapat korelasi antara sentimen komentar di Instagram dengan jumlah penjualan produk D'Alba di marketplace.

3.4.2 Variabel Data

Pada penelitian ini terdapat dua jenis variabel data yaitu variabel independen (fitur) dan variabel dependen (hasil output). Berikut dibawah ini penjelasan lebih lengkap mengenai kedua variabel tersebut [49].

3.4.2.1 Variabel Independen

Variabel independen merupakan atribut atau fitur dari komentar yang akan digunakan sebagai masukan dalam model analisis sentimen. Variabel ini diperoleh setelah melakukan data *preprocessing* dan vektorisasi. Pada penelitian ini, variabel yang digunakan adalah:

1. Komentar pengguna Instagram yang sudah dalam bentuk numerik hasil vektorisasi.
2. Harga produk *skincare* D'Alba.
3. *Rating* produk *skincare* D'Alba.

3.4.2.2 Variabel Dependen

Variabel dependen merupakan variabel hasil *output* atau variabel yang akan menjadi target untuk dianalisis yaitu hasil klasifikasi sentimen dari komentar yang diperoleh setelah menerapkan algoritma SVM dan Naïve Bayes, berikut merupakan variabel dependen yang digunakan:

1. Positif, komentar menunjukkan opini positif terhadap produk D'Alba.
2. Negatif, komentar mengandung kritik atau tidak puas terhadap produk D'Alba.
3. Netral, komentar tidak bersifat informatif yang menunjukkan opini positif atau negatif.
4. Kolom quantity_sold (jumlah produk yang terjual) pada data *marketplace*.

3.4.3 Teknik Pengambilan Sampel

Sampel yang digunakan adalah komentar yang ada di Instagram dalam rentang waktu tiga tahun yaitu 1 Januari 2021 sampai dengan 1 Desember 2023 dengan tagar #dAlbaIndonesia. Metode yang digunakan dalam penelitian ini

adalah *purposive sampling* [46], karena sampel yang diambil hanya komentar dalam rentang waktu tertentu dan yang membahas D'Alba berdasarkan tagar yang digunakan yaitu #dAlbaIndonesia.

3.5 Teknik Pengujian

Sebagai bagian dari analisis data berbasis *machine learning*, penelitian ini menggunakan pendekatan kuantitatif melalui pengujian hipotesis statistik. Teknik pengujian yang digunakan dalam penelitian ini terdiri dari dua bagian, yaitu evaluasi performa algoritma klasifikasi dan pengujian statistik hubungan sentimen dengan penjualan.

1. Evaluasi Performa Algoritma

Evaluasi performa dilakukan dengan menggunakan metode hold-out validation, yaitu membagi data ulasan menjadi 80% data latih (*training*) dan 20% data uji (*testing*). Dua algoritma yang digunakan adalah SVM dan Naïve Bayes. Model dilatih untuk mengklasifikasikan komentar menjadi tiga kategori sentimen: positif, negatif, dan netral. Kinerja kedua model dievaluasi menggunakan *confusion matrix*, evaluasi tersebut bertujuan untuk mengetahui algoritma mana yang memiliki performa lebih baik dalam analisis sentimen.

2. Uji Statistik (Uji Signifikansi p-value)

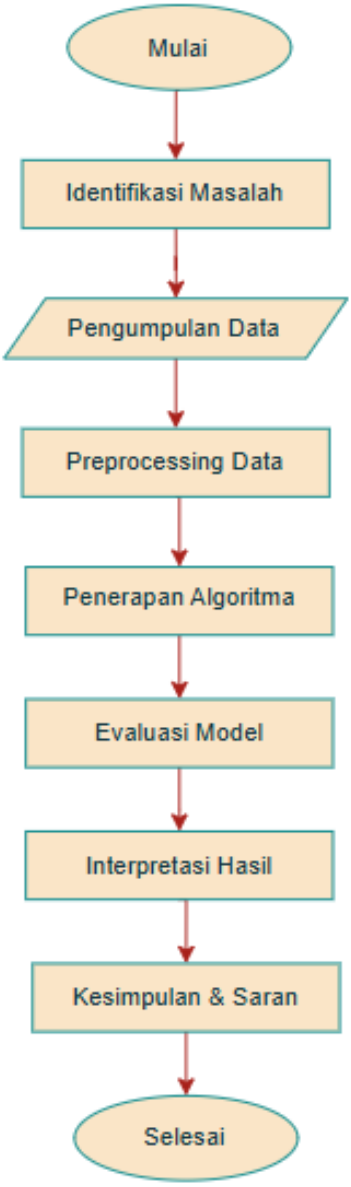
Untuk menguji apakah terdapat hubungan yang signifikan antara hasil klasifikasi sentimen dengan data penjualan produk di marketplace, digunakan analisis korelasi Spearman.Selanjutnya dilakukan pengujian signifikansi menggunakan uji p-value, di mana:

- H_0 (Hipotesis Nol): tidak terdapat hubungan yang signifikan antara sentimen dan penjualan.
- H_1 (Hipotesis Alternatif): terdapat hubungan yang signifikan antara sentimen dan penjualan.

Nilai p-value < 0,05 dianggap signifikan dan menjadi dasar untuk menolak hipotesis nol. Hasil uji ini membantu menyimpulkan apakah opini di media sosial memiliki pengaruh terhadap performa penjualan produk.

3.6 Alur Penelitian

Berikut Gambar 3.1 merupakan alur penelitian yang sesuai dengan framework CRISP-DM, flowchart disajikan dalam bentuk diagram dengan dimodifikasi namun tetap sesuai dengan CRISP-DM. Pembuatan alur penelitian ini menggunakan *draw.io* karena situsnya yang mudah digunakan, untuk penjelasan lebih lanjut dapat dilihat dibawah ini.



Gambar 3.1 Flowchart Alur Penelitian

3.6.1 Identifikasi Masalah (*Business Understanding*)

Penelitian ini dilakukan karena melihat gambaran tentang industri *skincare* yang berkembang pesat sesuai dengan data di latar belakang penelitian ini. Produk *skincare* D'Alba yang banyak diperbincangkan di media sosial, terutama Instagram, namun masih belum diketahui apakah sentimen konsumen terhadap *skincare* ini lebih dominan positif, negatif, atau bahkan netral. Tujuan dari penelitian ini adalah untuk menganalisis sentimen konsumen terhadap *skincare* D'Alba di Instagram serta membandingkan hasil akurasi mana yang lebih baik antara algoritma SVM dan Naïve Bayes dalam melakukan klasifikasi sentimen.

3.6.2 Pengumpulan Data (*Data Understanding*)

Data yang digunakan pada penelitian ini adalah komentar atau ulasan terkait produk D'Alba yang diperoleh dari Instagram selama periode tiga tahun yaitu 1 Januari 2021 hingga 1 Desember 2023, selama periode 3 tahun tersebut telah didapat kurang lebih 234 unggahan di Instagram mengenai D'Alba dengan tagar #dAlbaIndonesia dengan hasil sebanyak 3341 komentar untuk data latih.

Sedangkan untuk data uji diambil data pada periode 1 tahun yaitu 1 Januari 2024 hingga 1 Januari 2025 (kurang lebih 80 unggahan) dengan hasil sebanyak 1940 komentar. Total komentar yang didapat adalah 5281, hasil dari *scraping* data instagram berisikan beberapa kolom yaitu:

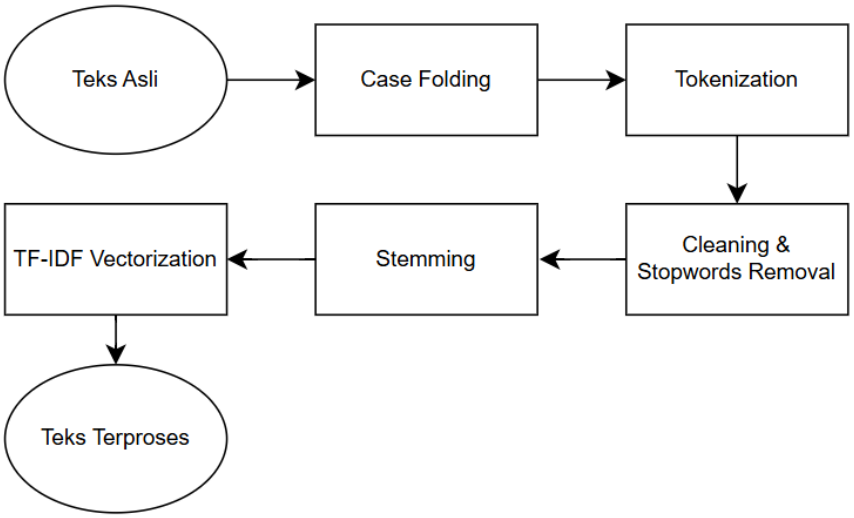
1. *username*: nama akun dari pengguna yang memberikan komentar.
2. *profile_id*: ID unik dari akun pengguna yang berkomentar.
3. *date*: tanggal beserta waktu komentar tersebut diunggah.
4. *comment*: isi komentar yang akan dianalisis sentimennya.
5. *source*: sumber data, berasal dari Instagram dan ketika di klik *link* akan langsung menuju ke komentarnya.

Kemudian untuk hasil dari *scraping* data *Marketplace* terdapat 798 baris yang berisikan beberapa kolom yaitu:

1. nama_produk: merupakan nama produk yang dijual di *marketplace*.
2. harga_produk: harga dari produk tersebut berbentuk Rupiah.
3. *rating*_produk: *rating* dari produk tersebut.

4. *quantity_sold*: jumlah produk yang terjual.

3.6.3 Pra-proses Data (*Data Preparation*)



Gambar 3.2 Alur *Preprocessing* Data

Pada tahap ini dilakukan pembersihan dan persiapan data komentar dari Instagram agar siap digunakan dalam analisis sentimen. Proses ini sangat penting karena data mentah seringkali mengandung noise seperti karakter khusus, emoji, atau kata-kata yang tidak relevan. Berikut merupakan tahapan *preprocessing* yang dilakukan dalam penelitian ini [47]:

1. ***Case Folding***

   Mengubah semua teks menjadi huruf kecil untuk menseragamkan format data, dengan melakukan proses ini dapat meningkatkan akurasi model karena model tidak akan salah menganggap kata yang sama dalam bentuk huruf berbeda. Berikut contohnya:

Tabel 3.6 Contoh Case Folding

Sebelum Case Folding	Sesudah Case Folding
White truffle double serum & cream bagus bangett 🔥🔥🔥	white truffle double serum & cream bagus bangett 🔥🔥🔥
Produk yg Nikita Willy dengar pertama kali dari MUAnya adalah d'Alba First Spray Serum yang bisa menghidrasi kulit saat make up agar tidak cakey dan lebih flawless😍 @dalba_indonesia	produk yg nikita willy dengar pertama kali dari muanya adalah d'alba first spray serum yang bisa menghidrasi kulit saat make up agar tidak cakey dan lebih flawless😍 @dalba_indonesia

2. *Tokenization*

Merupakan proses membagi teks menjadi unit yang lebih kecil atau bisa disebut token, token sendiri berupa kata, karakter, atau sub-kata. Tujuan dilakukannya proses ini adalah agar dapat diolah lebih mudah oleh model *machine learning* dan mengurangi noise (karakter tidak penting). Berikut contohnya:

Tabel 3.7 Contoh *Tokenization*

Sebelum *Tokenization*	Sesudah *Tokenization*
white truffle double serum & cream bagus bangett 🔥 🔥 🔥	[“white”, “truffle”, “double”, “serum”, “&“, “cream”, “bagus”, “bangett”, “ 🔥 🔥 🔥 ”]
produk yg nikita willy dengar pertama kali dari muanya adalah d'alba first spray serum yang bisa menghidrasi kulit saat make up agar tidak cakey dan lebih flawless 😍 @dalba_indonesia	[“produk”, “yg”, “nikita”, “willy”, “dengar”, “pertama”, “kali”, “dari”, “muanya”, “adalah”, “d'alba”, “first”, “spray”, “serum”, “yang”, “bisa”, “menghidrasi”, “kulit”, “saat”, “make”, “up”, “agar”, “tidak”, “cakey”, “dan”, “lebih”, “flawless”, “ 😍 ”, “@dalba_indonesia”]

3. *Cleaning* dan *Stopword removal*

Proses ini bertujuan menghapus kata-kata umum yang tidak memiliki magna secara spesifik seperti “dan”, “atau”, “di”, “yang”, “ke” dan sebagainya dalam bahasa Indonesia. Kata-kata ini tidak memberikan kontribusi dalam pemrosesan teks, sehingga dapat dihapus agar bisa meningkatkan efisiensi dan akurasi model. Berikut contoh penggunaan *stopword removal*:

Tabel 3.8 Contoh *Stopword removal*

Sebelum *Stopword removal*	Sesudah *Stopword removal*
[“white”, “truffle”, “double”, “serum”, “&“, “cream”, “bagus”, “bangett”, “ 🔥 🔥 🔥 ”]	white truffle double serum cream bagus bangett
[“produk”, “yg”, “nikita”, “willy”, “dengar”, “pertama”, “kali”, “dari”, “muanya”, “adalah”, “d'alba”, “first”, “spray”, “serum”, “yang”, “bisa”, “menghidrasi”, “kulit”, “saat”, “make”, “up”, “agar”, “tidak”, “cakey”, “dan”,	produk nikita willy dengar pertama kali muanya d'alba spray serum bisa menghidrasi kulit make up cakey flawless

Sebelum *Stopword removal*	Sesudah *Stopword removal*
“lebih”, “flawless”, “😍”, “@dalba_indonesia”]	

4. *Stemming*

Stemming merupakan pemrosesan bahasa alami yang digunakan untuk mengubah kata ke bentuk dasar atau bentuk akarnya (stem), tujuannya agar bisa mengurangi variasi kata yang berasal dari kata yang sama. Seperti kata “berjalan”, “pejalan”, “jalan-jalan” akan diubah menjadi bentuk dasar yang sama yaitu “jalan”. Berikut contohnya:

Tabel 3.9 Contoh *Stemming*

Sebelum *Stemming*	Sesudah *Stemming*
white truffle double serum cream bagus bangett	white truffle double serum cream
produk nikita willy dengar pertama kali muanya d'alba spray serum bisa menghidrasi kulit make up cakey flawless	nikita willy dengar d’alba spray serum hidrasi kulit cakey flawless

5. *TF-IDF Vectorization*

TF-IDF atau Term Frequency-Inverse Document Frequency merupakan metode representasi teks ke bentuk vektor numerik untuk menilai seberapa penting suatu kata. Pada penelitian ini, komentar mengenai *skincare* D’Alba di Instagram diubah menjadi bentuk yang dapat digunakan oleh algoritma SVM dan Naïve Bayes. Berikut contohnya:

Tabel 3.10 Contoh Vektorisasi dengan TF-IDF

Kata Penting	Skor TF-IDF (Komentar 1)	Skor TF-IDF (Komentar 2)
bagus	0.45	0.00
serum	0.30	0.25
hidrasi	0.00	0.40
flawless	0.00	0.35

Kesimpulannya TF-IDF adalah mengidentifikasi kata penting dalam komentar, karena perkataan seperti “bagus”, “hidrasi”, “flawless” sering muncul sehingga dapat cenderung dikategorikan positif. Skor TF-IDF dihitung berdasarkan dua komponen utama yaitu Term Frequency yaitu

seberapa sering kata tersebut muncul, dan Inverse Document Frequency yaitu seberapa unik kata tersebut. Berikut merupakan rumus TF-IDF [37]:

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

Misal komentar untuk menghitung TF, "white truffle double serum cream bagus bangett", jumlah komentar adalah 6 dengan kata "bagus" adalah 1 maka nilai TF menjadi 1/6 yang hasilnya adalah 0.1167. Kemudian untuk menghitung IDF, contoh jika ada 100 komentar dengan 20 kata bagus maka menjadi log 100/20 sehingga menjadi log 5 dengan hasil menjadi 0.698. Hasil akhir menjadi TF-IDF yaitu 0.1667 x 0.698 = 0.1165.

3.6.4 Penerapan Algoritma (*Modeling*)

Pada penelitian ini digunakan dua algoritma klasifikasi yaitu SVM dan Naïve Bayes dalam proses melakukan analisis sentimen *skincare* D'Alba melalui komentar di Instagram D'Alba. Berikut tahapan dalam melakukan modeling:

1. ***Splitting Data***

   Proses ini dilakukan ketika tahap *preprocessing* sudah dilewati, terdapat data *train* yang digunakan untuk melatih model dan data *testing* yang digunakan untuk mengukur kinerja model tersebut. Umumnya rasio yang digunakan adalah 80% data *train*, dan 20% data *testing*. Hal ini dilakukan agar sebagian besar data digunakan untuk melatih model sehingga model dapat mengenali pola dari data secara optimal. Pendekatan tersebut cukup seimbang karena memberikan cukup banyak data bagi model untuk belajar dan tetap menyediakan data (20%) untuk mengukur kemampuan model secara adil [48].

2. **TF-IDF**

   Selanjutnya sebelum lanjut ke tahap modeling, data akan diubah menjadi bentuk numerik agar bisa diolah oleh algoritma. Dengan penggunaan TF-IDF vektorisasi untuk mengkonversi segala komentar

menjadi bentuk numerik agar bisa di proses oleh algoritma. Setiap kata komentar akan memiliki nilai bobot berdasarkan frekuensinya dan seberapa unik kata tersebut.

3. **Penerapan Algoritma**

   Terakhir, setelah data sudah menjadi bentuk numerik maka model akan dilatih menggunakan dua algoritma yaitu SVM dan Naïve Bayes. Untuk SVM sendiri akan mencari hyperplane terbaik dengan memisahkan data berdasarkan sentimen positif, negatif, dan netral. Berbeda dengan cara kerja Naïve Bayes yaitu menggunakan teorema Bayes untuk menghitung probabilitas suatu komentar, model ini akan mengasumsikan setiap kata dalam teks bersifat independen.

3.6.5 Evaluasi Model (*Evaluation*)

Setelah model dilatih dan diterapkan algoritmanya, maka akan dievaluasi dengan menggunakan beberapa metrik sebagai berikut:

1. Akurasi, merupakan hasil seberapa banyak prediksi yang benar dibandingkan dengan total data *testing*.
2. *Precision*, merupakan hasil seberapa banyak prediksi positif yang benar dari total yang diprediksi positif.
3. *Recall*, merupakan hasil seberapa banyak data positif yang berhasil diklasifikasikan dengan benar.
4. F-1 *Score*, merupakan hasil rata-rata dari antara *precision* dan *recall* untuk mengukur keseimbangan antara keduanya.

3.6.6 Interpretasi Hasil dan Kesimpulan (*Deployment*)

Setelah melakukan modeling dengan algoritma SVM dan Naïve Bayes, maka hal selanjutnya yang akan dilakukan adalah menganalisis hasil dari kedua algoritma tersebut dan menyimpulkan hasil temuan. Pada interpretasi hasil, model yang telah dievaluasi dengan metrik akan dibandingkan dan jika SVM memiliki akurasi lebih tinggi maka algoritma tersebut lebih efektif digunakan, namun sebaliknya jika Naïve Bayes memiliki *recall* lebih tinggi maka berarti algoritma ini lebih baik dalam menangkap pola sentimen.

Kemudian hasil akan berbentuk dalam laporan dan *website* yang menyajikan gambaran mengenai persepsi konsumen terhadap *skincare* D'Alba. Misal, jika mayoritas berkomentar sentimen positif berarti produk diterima dengan baik oleh konsumen, sebaliknya jika mayoritas berkomentar sentimen negatif berarti produk diterima kurang baik oleh konsumen dan perlu kritikan, atau jika sentimen dari konsumen bersifat netral maka berarti komentar tidak memberikan opini yang jelas atau hanya sekedar mention produk menggunakan emoji. Hasil analisis dapat dimanfaatkan oleh brand untuk meningkatkan strategi pemasaran dan kualitas produk berdasarkan masukan dari komentar konsumen.