BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Objek penelitian pada skripsi ini adalah ulasan masyarakat terhadap aplikasi belanja online (*e-commerce*) Shopee yang diperoleh melalui *review* yang diberikan oleh pengguna terhadap aplikasi Shopee pada *Google Play Store*. Media *Google Play Store* dipilih sebagai sumber data ulasan aplikasi Shopee dikarenakan, pada *Google Play Store* memungkinkan penggunanya untuk memberikan opini yang berkaitan dengan aplikasi yang digunakan [50]. Dari ulasan-ulasan yang diberikan oleh pengguna aplikasi tersebut, dapat diperoleh informasi-informasi yang berkaitan dengan fitur-fitur yang perlu untuk dikembangkan atau dipertahankan oleh pengembang aplikasi dalam memberikan performa aplikasi yang lebih baik kedepannya. Untuk mempermudah pengelolaan data ulasan yang diberikan oleh pengguna aplikasi Shopee, maka diperlukan analisis yang mendalam terhadap ulasan-ulasan yang diberikan oleh pengguna aplikasi Shopee yang terdapat pada *Google Play Store*.

Data ulasan yang akan digunakan dalam penelitian ini akan dikumpulkan dalam rentang waktu antara 1 Januari 2023 hingga 31 Desember 2023, mengingat periode tersebut merupakan puncak popularitas aplikasi Shopee tertinggi jika dibandingkan dengan aplikasi *e-commerce* lainnya dengan jumlah pengguna mencapai 2,35 juta pengunjung selama tahun 2023 [51]. Data yang akan digunakan dalam penelitian yaitu sebanyak 23.884 baris data dengan 10 kolom, yaitu kolom *reviewId, userName, userImage, content, score, thumbsUpCount, reviewCreatedVersion, at, replyContent,* dan *repliedAt*. Data ulasan yang akan digunakan pada penelitian ini akan dibagi menjadi dua kategori, yaitu kategori positif dan kategori negatif. Pembagian data menjadi dua kategori tersebut berdasarkan penelitian yang telah dilakukan sebelumnya dalam penelitian [12],

[13] dan [14], tingkat akurasi yang dicapai menunjukkan tingkat akurasi yang cukup baik. Pada tahap selanjutnya, data sentimen masyarakat tersebut akan dilakukan analisis sentimen dengan menggunakan algoritma *Support Vector Machine* (SVM) dan algoritma *K Nearst Neighbors* (KNN) untuk memperoleh hasil *accuracy*, *precision*, *recall*, dan *f1-score.*

Setelah mendapatkan hasil dari model menggunakan algoritma *Support Vector Machine* (SVM) dan KNN, langkah berikutnya adalah melakukan optimasi model. Tujuan dari optimasi model adalah untuk meningkatkan hasil klasifikasi sentimen pengguna aplikasi Shopee. Proses optimasi model akan menerapkan teknik *GridSearch CV,* yang melibatkan penyesuaian parameter tertentu dalam algoritma seperti fungsi *kernel* untuk SVM atau jumlah tetangga terdekat untuk algoritma KNN. Tujuan dari optimasi ini adalah untuk meningkatkan kinerja model secara keseluruhan. Selain itu, teknik *cross-validation* juga akan diterapkan untuk memastikan bahwa model tidak hanya efektif pada data pelatihan tetapi juga mampu generalisasi dengan baik pada data baru. Langkah-langkah ini krusial untuk memastikan bahwa model yang dihasilkan tidak mengalami *overfitting* atau *underfitting* terhadap data.

Penelitian ini bertujuan untuk menyusun klasifikasi terhadap komentar pengguna di Google Play Store terkait aplikasi Shopee, dengan tujuan memberikan rekomendasi yang dapat meningkatkan pengembangan aplikasi Shopee di masa mendatang.

3.2 Metode Penelitian

3.2.1 Metode Pengembangan Sistem Data Mining

Dalam penelitian ini, digunakan metodologi *data understanding methodology*. Pada tabel 3.1 menunjukkan perbandingan metodologi yang sering digunakan yaitu: *Cross Industry Standard Process for Data Mining* (CRISP-DM), *Knowledge Discovery in Databases* (KDD), dan *Sample, Explore, Modify, Model, Assess* (SEMMA).

Tabel 3.1 Tabel perbandingan CRISP-DM, KDD dan SEMMA

Indikator	**CRISP-DM**	**KDD**	**SEMMA**
Tahapan	*1.Bussiness understanding* *2.Data understanding* *3. Data preparation* *4. Modelling* *5. Evaluation* *6. Deployment*	*1. Business understanding* *2. Data selection* *3. Data preprosessing* *4. Data transformation* *5. Data Mining* *6. Evaluation* *7. Interpretation*	*1. Sample data* *2. Explore data* *3. Modify* *4. Model* *5. Assesment*
Kesimpulan	Metodologi CRISP-DM menekankan pada siklus iteratif yang memungkinkan untuk penyesuaian dan perbaikan berkelanjutan.	Metodologi KDD lebih fleksibel dan memungkinkan integrasi dengan berbagai metodologi dan alat analisis data.	Metodologi SEMMA berfokus pada analisis data dan pemodelan, sering digunakan dalam lingkungan yang menggunakan perangkat lunak SAS.

Berdasarkan tabel 3.1 diatas, pada penelitian ini, akan mengadopsi metodologi CRISP-DM (*Cross Industry Standard Process for Data Mining*) sebagai pendekatan utama dalam penelitian. Pilihan ini didasarkan pada perbandingan antara tiga metodologi utama pada tabel 3.1, yaitu CRISP-DM, KDD, dan SEMMA. Metodologi CRISP-DM dipilih karena menawarkan kerangka kerja yang terstruktur dengan enam tahap, mencakup *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment* [52]. Hal tersebut jika dibandingkan dengan metodologi KDD yang bersifat lebih luas dan metodologi SEMMA yang lebih fokus pada *data analysis*, CRISP-DM memberikan keseimbangan yang baik antara kerangka kerja yang terperinci dan fleksibilitas [52]. Oleh karena itu, metodologi CRISP-DM diharapkan memberikan arahan yang jelas dalam menjalankan proses *data mining* serta memfasilitasi pengambilan keputusan yang informasional dan efektif.

Berikut merupakan penjelasan terhadap tahapan-tahapan yang terdapat pada metodologi CRISP-DM:

3.2.1.1 Business Understanding

Pada tahap pertama yaitu tahap *business understanding*, dalam metodologi CRISP-DM, tahap *business understanding* bertujuan untuk memahami tujuan dan konteks dari proyek penelitian yang akan dilakukan [53]. Dalam konteks pada penelitian, tujuan dilakukannya penelitian adalah untuk memahami sentimen yang dihasilkan oleh pengguna terhadap aplikasi Shopee selama menggunakan aplikasi Shopee pada *platform Google Play Store*. Tahap ini memungkinkan peneliti untuk menetapkan arah dan sasaran yang jelas, serta memastikan bahwa semua langkah yang diambil dalam proses penelitian berkontribusi secara langsung terhadap tujuan yang telah ditetapkan. Dengan memahami konteks bisnis dan tujuan yang jelas, peneliti dapat merancang dan menjalankan proses penelitian yang tepat dan relevan untuk mencapai hasil yang diinginkan.

3.2.1.2 Data Understanding

Data Understanding merupakan langkah yang diperlukan untuk memperoleh pemahaman yang komprehensif terhadap *dataset* yang digunakan pada penelitian, pada tahap *data understanding* mencakup proses pengumpulan data, eksplorasi data, mendeskripsikan isi data, dan melakukan evaluasi kualitas data [54]. Pada penelitian ini menggunakan data dari ulasan pengguna pada *Google Play Store* terhadap aplikasi Shopee. Data diperoleh melalui penggunaan *library google-play-scraper* dan data difilter berdasarkan rentang waktu dari 1 Januari 2023 hingga 31 Desember 2023. Meskipun dataset memiliki beberapa kolom ulasan dari *Google Play Store*, fokus penelitian ini adalah pada kolom *content* yang berisi ulasan pengguna, serta kolom *score* yang merupakan *rating* pengguna terhadap aplikasi Shopee, yang akan digunakan untuk menentukan sentimen positif dan negatif.

3.2.1.3 Data Preparation

Data preparation merupakan langkah penting dalam metodologi CRISP-DM, pada tahap *data preparation* akan mempersiapkan data agar siap diproses pada tahapan *data modeling* [55]. Pada tahap *data preparation* akan melakukan beberapa tahap beberapa diantaranya yaitu, *data cleaning* yang bertujuan untuk memastikan konsistensi format dan kebersihan data, seperti mengubah huruf kapital menjadi huruf kecil, menghapus angka, emoji, menghapus baris kosong pada dataset, menghapus baris duplikat, menghapus tanda baca, menghapus spasi kosong atau *whitespace*, dan menghapus beberapa kolom yang tidak digunakan pada penelitian. Kolom-kolom tersebut dihapus dikarenakan tidak digunakan pada penelitian kolom tersebut antara lain *reviewId*, *userName*, *userImage*, *thumbsUpCount*, *reviewCreatedVersion*, *replyContent*, dan kolom *repliedAt*. Pada tahap *data preparation* juga akan memberikan label pada data ulasan aplikasi Shopee dengan rentang *rating* 1 hingga *rating 5*. Pemberian label sentimen tersebut berdasarkan penelitian yang telah dilakukan oleh Oktian Permana [14] Sri Watmah [12], dan Muhammad Farid El Firdaus [13], pemberian *rating* dilakukan secara otomatis dengan bantuan bahasa pemrograman Python. Pemberian *labeling* terhadap dataset akan memanfaatkan *library transformers* pada python. Ulasan dengan *rating score* 3 tidak digunakan pada penelitian ini, dikarenakan *rating* 3 memiliki *content review* yang ambigu dan tidak merepresentasikan *label* dari Negatif maupun *label* dari Positif [13].

3.2.1.4 *Data Modeling*

Pada tahap *data modelling* dalam metodologi CRISP-DM *(Cross-Industry Standard Process for Data Mining*), data yang telah melewati tahap *data preparation* pada tahap sebelumnya akan dieksekusi dalam pengembangan model. Pada tahap ini dimulai dengan pembagian data menjadi dua bagian, yaitu *data training* dan *data testing*. Untuk memudahakan dalam membagi dataset menjadi *data training* dan *data* testing, menggunakan *library* dari Python yaitu *train_test_split*. *Data*

training digunakan untuk melatih model terhadap pola dari label sentimen yang telah diberikan, sedangkan *data testing* digunakan untuk menguji performa model yang telah dibentuk.

Pada tahap *data modelling* dilakukan *feature extraction* dengan menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*). Metode TF-IDF bertujuan untuk memberikan bobot pada setiap kata berdasarkan tingkat kemunculannya dalam dokumen dan invers frekuensi dokumen di seluruh kumpulan dokumen. Dengan demikian, kata-kata yang muncul secara sering dalam satu dokumen tetapi jarang muncul dalam kumpulan dokumen akan memiliki bobot yang lebih tinggi [56].

Setelah proses *feature extraction* selesai, langkah selanjutnya adalah pembuatan model menggunakan dua algoritma, yaitu *Support Vector Machine* (SVM) dan *K-Nearest Neighbors* (KNN). Algoritma SVM merupakan algoritma pembelajaran yang digunakan untuk klasifikasi atau regresi, yang bertujuan untuk menemukan *hyperplane* terbaik yang memisahkan antara dua kelas data sedangkan algoritma KNN adalah algoritma pembelajaran yang berbasis *instance*, di mana prediksi dilakukan dengan cara mencari kelas mayoritas dari k tetangga terdekat dari titik data yang akan diprediksi.

Dengan menggabungkan proses pembagian data, ekstraksi fitur menggunakan TF-IDF, dan penerapan dua algoritma pembelajaran mesin yang berbeda, diharapkan model yang dihasilkan dapat memberikan prediksi yang akurat dan dapat diandalkan untuk masalah yang sedang dihadapi. Selain itu, proses ini memungkinkan untuk melakukan pembandingan kinerja antara dua algoritma yang berbeda dalam menyelesaikan tugas yang sama, sehingga memungkinkan untuk memilih model yang paling sesuai dengan kebutuhan dan karakteristik data yang dimiliki.

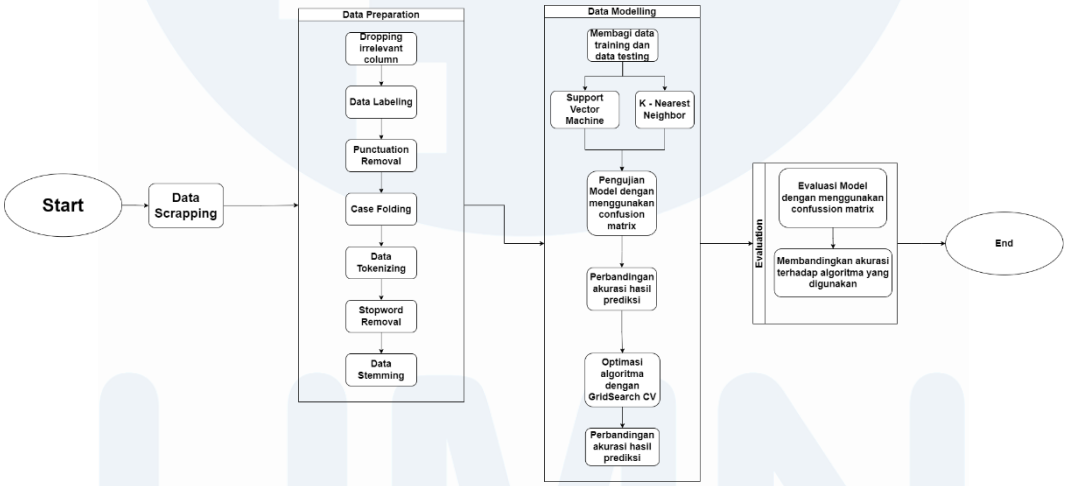
3.2.1.5 Evaluation

Tahapan evaluasi merupakan langkah penting dalam proses pengembangan model setelah tahap pemodelan data. Pada tahap ini, performa model-model yang telah dibentuk akan dievaluasi secara mendalam untuk memahami seberapa baik model-model tersebut bekerja dalam mengklasifikasikan sentimen pengguna Shopee. Proses evaluasi ini menggunakan metrik-metrik seperti *confusion matrix*. *Confusion matrix* memberikan gambaran yang jelas tentang seberapa baik model dapat mengklasifikasikan data ke dalam kelas yang benar, serta seberapa sering model melakukan kesalahan dalam klasifikasi. Dengan *confusion matrix* memungkinkan untuk mengetahui akurasi dari masing-masing model dengan memperhitungkan hasil dari *precision, recall, f1-score, dan accuracy*.

Setelah hasil model pada algoritma *Support Vector Machine* dan KNN diperoleh, selanjutnya model yang dibentuk akan dilakukan optimasi. Hal tersebut untuk memberikan hasil yang lebih baik dalam mengklasifikasikan sentimen pengguna Shopee. Proses optimasi model akan menggunakan teknik *GridSearch CV*. Pada teknik *GridSearch CV* melibatkan penyesuaian parameter-parameter tertentu dalam algoritma, seperti kernel function untuk SVM atau jumlah tetangga terdekat untuk KNN, sehingga meningkatkan performa model secara keseluruhan. Selain itu, teknik seperti *cross-validation* juga dapat diterapkan untuk memastikan bahwa model tidak hanya bekerja dengan baik pada data pelatihan tetapi juga mampu menggeneralisasi dengan baik pada data baru. Langkah-langkah ini penting untuk memastikan bahwa model yang dihasilkan tidak *overfitting* atau *underfitting* terhadap data. Dengan demikian, tahapan evaluasi dan optimasi merupakan bagian integral dari proses pengembangan model yang memastikan model yang dihasilkan dapat memberikan hasil yang akurat dan dapat diandalkan dalam mengklasifikasikan sentimen pengguna Shopee secara efektif.

Setelah perbandingan antara algoritma-algoritma yang berbeda selesai, langkah selanjutnya adalah melakukan visualisasi data. Visualisasi data memiliki peran penting dalam proses analisis data, karena memungkinkan untuk memahami pola-pola yang tersembunyi dan hubungan-hubungan yang kompleks dalam data. Dalam konteks evaluasi model visualisasi *bar chart*, untuk menggambarkan kata-kata yang menjadi bagian dari sentimen positif dan sentimen negatif terhadap aplikasi Shopee. Melalui visualisasi ini, dapat dengan lebih mudah memahami pola-pola yang muncul dalam sentimen pengguna terhadap aplikasi, serta melakukan analisis lebih lanjut untuk mendukung pengambilan keputusan yang lebih baik.

3.2.2 Alur Penelitian



Gambar 3.1 Diagram Alur Penelitian

Pada gambar 3.1 merupakan diagram alur penelitian yang akan dilakukan pada penelitian. Berikut merupakan penejelasan mengenai alur penelitian tersebut:

3.2.2.1 Data scrapping

Awalnya, penelitian dimulai dengan pengumpulan data *rating* dan *review* pengguna aplikasi Shoppe pada *platform Google Play Store*. Digunakan teknik *web scraping*, yang merupakan metode otomatis untuk mengekstraksi data dari halaman website ke berbagai format yang dapat diolah untuk analisa, pada penelitian ini akan menggunakan file dengan format CSV (*Comma Separated Values*) [57]. Tahap *web scraping* pada

penelitian ini akan memanfaatkan bahasa pemrograman *python* dengan *library google_play_scraper* yang mana *library* ini telah diprogram untuk menjelajahi halaman *google play store* dan menemukan informasi yang relevan untuk digunakan pada penelitian seperti *rating*, *review* aplikasi, versi aplikasi, tanggal pengguna memberikan *review*, dan tanggapan *developer* aplikasi terhadap *review* yang diberikan oleh pengguna, data-data hasil *web scrapping* tersebut akan disimpan ke penyimpanan *local* untuk memudahkan dalam melakukan analisis sentimen. Dokumen hasil *web scrapping* akan disimpan dalam format CSV [58]. Teknik ini memungkinkan peneliti untuk mengumpulkan sejumlah besar data secara efisien dan akurat untuk analisis lebih lanjut dalam penelitian.

3.2.2.2 *Data Preparation*

Pada tahap *preparation data*, merupakan langkah penting dilakukan untuk mengubah informasi yang diperoleh dari sumber data yang telah di *scrapping* sebelumnya menjadi format baku atau struktur data yang lebih terorganisir. Hal ini penting karena data hasil *web scraping* pada dasarnya cenderung tidak terstruktur. Pada tahap *preprocessing* meliputi beberapa tahapan kompleks yaitu, termasuk memberikan *data labeling* pada dataset, membersihkan data (*cleaning*), membagi teks menjadi token (*tokenizing*), melakukan filtrasi data (*filtering*), dan mengekstrak kata menjadi bentuk dasarnya tanpa adanya imbuhan pada awal kata maupun akhir kata (*stemming*) [59]. Setelah melalui serangkaian langkah tersebut, hasilnya adalah data yang telah diolah secara rapi dan terstruktur, yang akan sangat memudahkan proses klasifikasi data yang akan dilakukan pada tahap-tahap selanjutnya. Dengan melakukan tahap *preprocessing* ini, informasi yang awalnya sulit dimengerti dan diolah dapat diubah menjadi bentuk yang lebih mudah dipahami dan dimanfaatkan dalam analisis dan pengambilan keputusan.

Pada gambar 3.2 dibawah, menunjukan alur tahapan *data preparation* pada penelitian.

Gambar 3.2 Tahapan data preparation

A. Dropping Irrelevant Column

Pada tahap *dropping irrelevant column* yang mana merupakan tahap untuk menghapus kolom pada dataset yang tidak digunakan dalam penelitian yakni kolom *reviewId, userName, userImage, thumbsUpCount, reviewCreatedVersion, replyContent, repliedAt,* dan *appVersion.*

B. Labeling

Tahap *labeling* merupakan tahap penting pada penelitian untuk memberikan label terhadap dataset yang didapatkan setelah tahap *data scrapping*. Pada tahap *labeling* setiap *review* pengguna akan diberikan *label* berdasarkan *rating* yang diberikan oleh pengguna aplikasi, pemberian label pada dataset memanfaatkan *library transformers* dari Python. *Library transformers* merupakan model *deep learning* yang dapat diguanakan untuk mengidentifikasi kalimat positif dan negatif. Setelah data *review* diberikan *label*, data tersebut dimasukan kedalam kolom baru pada dataset dengan nama kolom 'sentiment' untuk menampung hasil dari *labeling reivew* pengguna aplikasi Shopee. Data hasil *labeling* tersebut nantinya akan digunakan untuk melatih model dalam mengenal ulasan dengan label positif maupun negatif [60].

C. Punctuation Removal

Cleaning merupakan tahap pertama yang dilakukan pada tahap *preprocessing*. Tahap *cleaning* bertujuan untuk menghilangkan atau menghapus data yang bersifat tidak relevan atau dapat memberikan *noise* pada dataset [61]. Contoh dari tahapan *cleaning* adalah menghapus atau menghilangkan tanda baca pada kalimat, dengan tujuan untuk membuat data menjadi seragam. Pada tahap *cleaning* akan menghapus tanda baca, *emoji*, angka, dan link url, menghapus beberapa kolom yang

tidak diguankan dalam penelitian, menghapus *review* dengan *rating score* 3, melakukan *filtering* terhadap data dengan rentang tahun 2023. Contoh dari tahapan *cleaning* ditunjukan pada tabel 3.2 dibawah ini.

Tabel 3.2 Hasil tahapan Cleaning

Data Input	Data output (*Cleaning*)
Saya tidak masalah dengan shoppe nya**,** tapi SPX nya benar benar mengecewakan**.** Paket sudah sampe di kota tujuan**,** tapi sampai tiga hari tidak di antar**!** Jadi saya jemput sendiri**.** Pengiriman nya benar benar mengecewakan**!!!!!!!!** Tolong di perbaiki ya**.** Thx	saya tidak masalah dengan shoppe nya tapi spx nya benar benar mengecewakan paket sudah sampe di kota tujuan tapi sampai tiga hari tidak di antar jadi saya jemput sendiri pengiriman nya benar benar mengecewakan tolong di perbaiki ya thx

Pada tabel 3.2 tersebut, dapat diperhatikan terdapat beberapa karakter yang dihapus, seperti angka (**1,2,3**) dan tanda baca seperti '**!**', '**.**', '**,**'.

D. Case Folding

Pada tahap *case folding* bertujuan untuk mengubah data dengan huruf kapital menjadi huruf *lowercase*.

E. Data Tokenizing

Tahap *tokenizing* dalam analisis teks adalah proses penting yang melibatkan pembagian teks menjadi unit-unit yang lebih kecil, seperti kata-kata atau frasa, yang disebut dengan token [62]. Tujuan dari tahap *tokenizing* adalah untuk memecah teks menjadi bagian-bagian yang lebih kecil sehingga memudahkan dalam melakukan analisis data lebih lanjut [63]. Dengan demikian, setiap kata dapat diproses secara terpisah untuk melakukan analisis sentimen, penghitungan frekuensi kata, atau analisis lainnya. Tahap *tokenizing* penting karena memungkinkan untuk mengubah teks menjadi representasi yang lebih terstruktur dan siap untuk diolah lebih lanjut dengan alat-alat analisis teks, seperti teknik-teknik

pada *machine learning* atau analisis statistik [64]. Pada tabel 3.3 merupakan contoh data setelah melalui tahap *tokenizing*.

Tabel 3.3 Hasil tahapan tokenizing

Data Input	Data Output (Tokenizing)
Saya tidak masalah dengan shoppe nya tapi SPX nya benar mengecewakan Paket sudah sampe di kota tujuan tapi sampai hari tidak di antar Jadi saya jemput sendiri. Pengiriman nya benar mengecewakan Tolong di perbaiki ya Thx	["Saya", "tidak", "masalah", "dengan", "shoppe", "nya", "tapi", "SPX", "nya", "benar", "mengecewakan", "Paket", "sudah", "sampe", "di", "kota", "tujuan","tapi", "sampai", "hari", "tidak", "di", "antar", "Jadi", "saya", "jemput", "sendiri", "Pengiriman", "nya", "benar", "mengecewakan", "Tolong", "di", "perbaiki", "ya", "Thx"]

F. Stopwords Removal

Setelah tahap *tokenizing*, langkah selanjutnya adalah melakukan *stopword removal*, di mana pada tahap ini akan dilakukan penghapusan terhadap kata-kata yang umum digunakan dan tidak memiliki arti apapun [65]. Contoh kata-kata yang termasuk dalam kelompok *stopword* adalah "yang", "dan", "di", "itu", "dengan", "untuk", "tidak", "dari", "dalam", "akan", "pada", "ini", "juga", "saya", "serta", "adalah", "bahwa", "lain", "kamu", dan sebagainya. Pada tabel 3.4 merupakan contoh data setelah melalui tahap *stopword removal*.

Tabel 3.4 Hasil tahapan Stopwords Removal

Data Input	Data Output (Tokenizing)
[**"saya",** **"tidak"**, "masalah", **"dengan"**, "shoppe", **"nya", "tapi"**, "spx", **"nya"**, "benar", "mengecewakan", "paket", "sudah", "sampe", **"di"**, "kota", "tujuan", **"tapi"**, "sampai", "hari", **"tidak"**, **"di"**, "antar", "jadi", **"saya"**, "jemput", "sendiri", "pengiriman", **"nya"**, "benar", "mengecewakan", "tolong", **"di"**, "perbaiki", **"ya"**, "thx"]	["masalah", "shoppe", "spx", "benar", "mengecewakan", "paket", "sudah", "sampe", "kota", "tujuan", "sampai", "hari", "antar", "jadi", "jemput", "sendiri", "pengiriman", "benar", "mengecewakan", "tolong", "perbaiki", "thx"]

G. *Stemming*

Stemming adalah bagian dari tahap *preprocessing* teks yang bertujuan untuk mengubah kata-kata menjadi bentuk dasarnya tanpa adanya imbuhan [65]. Proses ini akan menghapus imbuhan seperti awalan dan akhiran sehingga hanya meninggalkan akar katanya saja dari bentuk infleksi, prefiks, sufiks, yang ada pada setiap kata. Pada tabel 3.5 menampilkan hasil data setelah melalui tahap *stemming*.

Tabel 3.5 Hasil tahapan Stemming

Data Input	Data Output (Tokenizing)
[“masalah”, “shoppe”, “spx”, “benar”, **“mengecewakan”**, “paket”, “sudah”, **“sampe”**, “kota”, “tujuan”, “sampai”, “hari”, “antar”, “jadi”, “jemput”, “sendiri”, **“pengiriman”**, “benar”, **“mengecewakan”**, “tolong”, **“perbaiki”**, “thx"]	["masalah", "shoppe", "spx", "benar", **"kecewa"**, "paket", "sudah", **"sampai",** "kota", "tujuan", "sampai", "hari", "antar", "jadi", "jemput", "sendiri", **"kirim”**, "benar", **"kecewa"**, "tolong", **"baik"**, "thx"]

3.2.2.3 Pemodelan data

Pada tahap data *modelling*, data akan dibagi menjadi dua bagian yaitu data *training* dan *data testing*. Pembagian data menjadi data *training* dan dat *testing* tersebut menggunakan *library* python yaitu *library train_test_split* data *training* digunakan untuk melatih data terhadap model yang dibentuk. Pada penelitian ini akan membagi *dataset* dengan *presentase* 70% untuk data *training* dan 30% untuk data *testing*. Pembagian data dengan rasio 70:30 tersebut berdasarkan penelitian yang telah dilakukan sebelumnya, dan pada penelitian tersebut memperoleh hasil akurasi model yang cukup tinggi [10] [20].

3.2.2.4 *Evaluation*

Setelah data melalui tahap *splitting data* menjadi dua bagian, langkah selanjutnya yaitu membuat *modeling* algoritma dengan menggunakan algoritma *Support Vector Machine* (SVM) dan *K Nearest Neighbors* (KNN). Algoritma tersebut akan dilatih terlebih dahulu dengan menggunakan data *training* yang telah dipersiapkan sebelumnya untuk memahami pola yang terdapat pada data dengan sentimen positif maupun dengan sentimen negatif. Setelah model memahiami pola yang terdapat pada data, maka model akan dilakukan *testing* dengan menggunakan data *testing*. Hasil akurasi dari algoritma tersebut akan dihitung berdasarkan *confusion matrix*. Hasil dari *confusion matrix* akan mencari nilai dari *accuracy*, *precision*, *recall*, dan *f1-score.*

Setelah mendapatkan hasil dari model menggunakan algoritma *Support Vector Machine* dan KNN, langkah selanjutnya adalah melakukan optimasi terhadap model. Tujuan dari optimasi ini adalah untuk meningkatkan kinerja model dalam mengklasifikasikan sentimen pengguna aplikasi Shopee. Proses optimasi akan mengimplementasikan teknik *GridSearch CV*, yang melibatkan penyesuaian parameter khusus dalam algoritma, seperti fungsi *kernel* untuk SVM atau jumlah *n_neighbors* untuk KNN. Tujuan dari tahap optimasi adalah untuk meningkatkan performa model secara keseluruhan. Selain itu, metode *cross-validation* juga akan digunakan untuk memastikan bahwa model tidak hanya efektif pada data pelatihan tetapi juga mampu menggeneralisasi dengan baik pada data baru. Langkah-langkah ini sangat penting untuk memastikan bahwa model yang dihasilkan tidak mengalami *overfitting* atau *underfitting* terhadap data. Keseluruhan proses evaluasi dan optimasi ini adalah bagian penting dalam pengembangan model, yang bertujuan untuk memberikan hasil yang akurat dan dapat diandalkan dalam mengklasifikasikan sentimen pengguna Shopee.

3.3 Teknik Pengumpulan Data

3.3.1 Populasi dan Sampel

Populasi yang akan digunakan dalam penelitian ini adalah ulasan yang diposting di *Google Play Store* terkait dengan aplikasi Shopee. Metode *sampling* yang digunakan adalah *non-probability* sampling yang dikenal sebagai *purposive sampling*. *Purposive sampling* merupakan pendekatan di mana dipilih sampel berdasarkan tujuan penelitian yang telah ditetapkan sebelumnya. Oleh karena itu, teknik ini dianggap memiliki keunggulan dalam mendapatkan data yang lebih akurat, serta menghasilkan hasil penelitian yang lebih optimal jika dibandingkan dengan teknik *sampling* lainnya [66]. Untuk mengambil data sampel, menggunakan *library* dari Python *google-play-scraper* melalui *Google Colaboratory* yang memungkinkan ekstraksi data menggunakan bahasa pemrograman *Python*.

3.3.2 Periode Pengambilan Data

Dalam konteks penelitian ini, dilakukan pengumpulan data ulasan terhadap aplikasi Shopee dari *platform* Google Play Store dalam rentang waktu, yakni dari tanggal 1 Januari 2023 hingga 31 Desember 2023. Penetapan periode ini disesuaikan dengan tujuan untuk memperoleh dataset yang representatif terhadap sentimen pengguna selama periode waktu yang cukup luas. Dengan meliputi satu tahun penuh dalam rentang waktu tersebut, diharapkan dapat mendapatkan data yang signifikan dalam jumlah dan variasi, yang kemudian akan menjadi dasar bagi analisis lanjutan dalam memahami dinamika respons pengguna terhadap aplikasi Shopee. Selain itu, pemilihan rentang waktu tertentu juga didasarkan pada pertimbangan untuk menjaga keseimbangan jumlah ulasan yang terkumpul setiap tahunnya, sehingga memungkinkan perbandingan yang objektif dan valid antar-tahun terkait dengan sentimen pengguna terhadap aplikasi Shopee.

Dari *Google Play Store*, berhasil terhimpun total sejumlah 23.884 ulasan yang terkait dengan aplikasi Shopee. Jumlah data yang besar ini memberikan kekayaan informasi yang substansial bagi penelitian ini, serta menjamin keandalan dan validitas analisis yang akan dilakukan. Dengan

dataset yang mencakup jumlah ulasan yang signifikan, penelitian ini akan memiliki dasar yang kuat untuk melakukan analisis yang mendalam terhadap faktor-faktor yang memengaruhi sentimen pengguna terhadap aplikasi Shopee dalam periode waktu yang telah ditentukan.

3.4 Variabel Penelitian

3.4.1 Variabel Independen

Variabel Independen merupakan variabel yang memberikan pengaruh atau perubahan terhadap variabel dependen. Variabel independen dari penelitian ini merupakan komentar/opini masyarakat terhadap aplikasi Shopee yang diperoleh melalui Google Play Store.

3.4.2 Variabel Dependen

Variabel dependen merupakan variabel yang memilliki hubungan keterkaitan dengan variabel independen atau nilai *value* dari variabel dependen dipengaruhi oleh variabel independen. Variabel dependen dari penelitian ini adalah sentimen masyarakat mengenai aplikasi Shopee, data dependen yang di pakai adalah sentimen negatif/sentimen positif.

3.5 Teknik Analisis Data

Penelitian ini mengadopsi pendekatan kualitatif karena data yang dikumpulkan berbentuk ulasan, yang kemudian dianalisis secara kualitatif. Pengolahan data dilakukan dengan menggunakan bahasa pemrograman Python, dengan memanfaatkan Google Colaboratory sebagai alat bantu. Data akan melalui tahap *pre-processing* terlebih dahulu, di mana dilakukan persiapan data untuk analisis, termasuk penghapusan *noise* dalam data. Selanjutnya, dilakukan *feature extraction* menggunakan metode TF-IDF (*term frequency-inverse document frequency*) untuk mengekstrak fitur-fitur penting dari ulasan. Analisis data dilakukan dengan menerapkan dua algoritma, yaitu Support Vector Machine (SVM) dan *K-Nearest Neighbors* (KNN), guna memperoleh pemahaman yang lebih mendalam tentang sentimen pengguna terhadap aplikasi Shopee. Akhirnya,

penelitian akan diakhiri dengan pengujian akurasi performa algoritma menggunakan *confusion matrix*.