BAB III

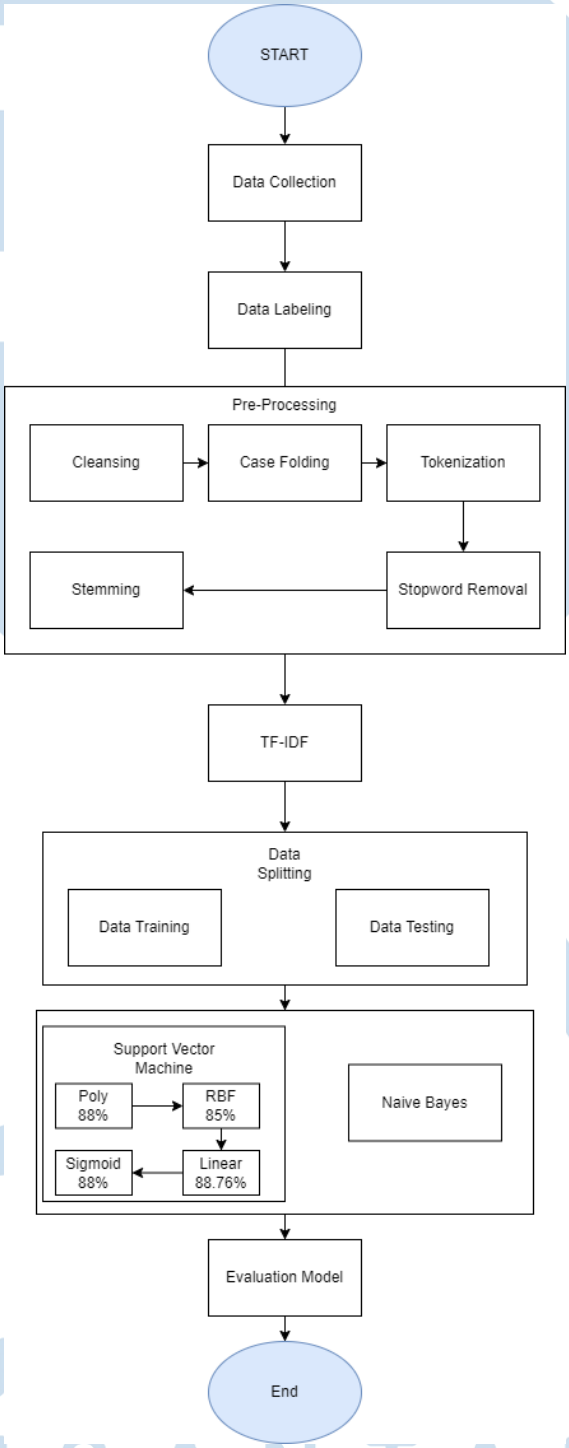
METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Objek penelitian ini adalah ulasan pengguna aplikasi TikTok yang tersedia di Google Playstore. TikTok, sebagai platform media sosial yang populer, memungkinkan penggunanya untuk berbagi video singkat dan telah menjadi fenomena global di kalangan generasi muda. Ulasan yang ditinggalkan oleh pengguna di Google Playstore merupakan sumber data yang berharga untuk memahami persepsi dan pendapat mereka terhadap aplikasi ini. Ulasan ini mencakup berbagai aspek, mulai dari fungsi aplikasi, pengalaman pengguna, hingga isu-isu teknis dan privasi. Data ulasan ini bersifat teks dan beragam, mencerminkan berbagai pengalaman dan opini pengguna. Mengingat popularitas dan dampak sosial TikTok, analisis sentimen terhadap ulasan ini dapat memberikan wawasan berharga mengenai kekuatan dan kelemahan aplikasi dari perspektif pengguna. Selain itu, analisis ini juga dapat membantu pengembang aplikasi untuk mengidentifikasi area-area yang memerlukan peningkatan.

3.2 Alur Penelitian

Alur penelitian pada Tiktok akan diawali dengan kerangka berpikir yang telah di sesuaikan dengan Metode KDD, pada Gambar 3.2:



Gambar 3. 1 Alur Penelitian

3.2.1 Data Collection

Selama tahap Pengumpulan Data, proses scrapping dilaksanakan menggunakan bahasa pemrograman *Python* di dalam lingkungan *Jupyter Notebook*. Skrip *Python* yang ditulis akan secara sistematis mengumpulkan data ulasan pengguna TikTok yang tersedia di Google Playstore. Proses ini akan mengambil informasi seperti teks ulasan, tanggal ulasan, jumlah bintang yang diberikan, dan mungkin informasi pengguna lainnya jika tersedia. Informasi ini kemudian akan disusun dan disimpan dalam format *Comma-Separated Values* (CSV), yang memudahkan analisis data lebih lanjut. Format CSV dipilih karena kompatibilitasnya yang luas dengan berbagai alat analisis data dan kemudahannya dalam manipulasi data.

3.2.2 Data Labeling

Dalam konteks penelitian ini, pelabelan data adalah proses penting di mana setiap entri data teks yaitu, ulasan pengguna diberi tag atau label sesuai dengan sentimen yang diwakilinya. Proses ini biasanya mengklasifikasikan teks menjadi kategori seperti positif, dan negatif. Penelitian ini menerapkan metode pelabelan otomatis menggunakan Python, di mana ulasan dengan rating 1-2 dikategorikan sebagai negatif, dan rating 4-5 sebagai positif. Setelah menggunakan pelabelan otomatis namun sering kali terjadi ketidaksesuaian antara isi komentar dan rating yang diberikan. Contoh ada pengguna yang memberikan komentar positif tetapi memberikan rating rendah atau memberikan komentar negatif tetapi memberikan rating tinggi. Untuk mengatasi hal ini, penulis melakukan pelabelan manual untuk memperbaiki dan memastikan kesesuaian antara komentar dan rating yang diberikan. Penulis memilih untuk melabeli data secara manual karena analisis sentimen untuk teks berbahasa Indonesia memerlukan pemahaman yang mendalam mengenai budaya, frasa idiomatik, dan kosakata lokal [42].

3.2.3 Pre-processing

Tahap ini menandai awal dari proses *text mining*, menjadi kunci dalam menyiapkan dokumen untuk analisis yang efektif. Pada tahap *pre-processing* ini

menggunakan Bahasa pemrograman Python, serangkaian langkah dilakukan untuk membersihkan dan menyempurnakan data, mulai dari pembersihan data guna menghilangkan unsur-unsur yang bisa mengganggu analisis, seperti noise dan informasi tidak relevan, sampai pada normalisasi teks untuk memastikan semua data memiliki format yang seragam.

1. *Cleansing*

   Data Cleaning adalah langkah awal dalam pra-pemrosesan data, di mana kesalahan, ketidakkonsistenan, dan ketidakakuratan dalam suatu set data diidentifikasi dan diperbaiki atau dihilangkan. Tujuan utama dari pembersihan data adalah untuk memastikan bahwa data cocok untuk keperluan analisis atau pemodelan. Proses ini melibatkan penanganan data yang hilang, mengeliminasi duplikasi, dan memperbaiki format atau struktur data. Pembersihan data menyiapkan dataset agar akurat, konsisten, dan siap untuk analisis atau pemodelan lebih lanjut. Penelitian ini menggunakan *library Regular Expressions*. *Regular Expressions* dapat mengidentifikasi dan mengubah pola teks yang spesifik. *Regex* dapat menghapus semua angka, URL, spasi ekstra, dan emoji dari teks untuk memastikan data yang bersih dan konsisten, yang siap untuk analisis lebih lanjut.

2. *Case folding*

   Langkah ini menyederhanakan dataset dengan mengonversi semua teks ke bentuk huruf kecil, mengurangi kompleksitas dan memfasilitasi pemrosesan data lebih lanjut dengan mengeliminasi perbedaan yang tidak signifikan antara huruf besar dan kecil.

3. *Tokenization*

   Proses ini melibatkan memecah teks menjadi token yang lebih kecil, seperti kata atau frase, menggunakan *library Natural Language Toolkit* (NLTK). Tujuan dari tokenisasi ini adalah untuk memfasilitasi analisis terperinci pada level kata atau frase, memungkinkan pemahaman yang lebih mendalam tentang struktur dan konten teks.

4. *Stopword removal*

Proses *stopword removal* menggunakan pustaka *Natural Language Toolkit* (NLTK) dan ditargetkan pada kata-kata umum dalam bahasa Indonesia, bertujuan untuk mengeliminasi kata-kata yang tidak menambah nilai analitis. Kata-kata seperti "dan", "atau", "tapi", dan lainnya sering muncul tetapi tidak memberikan informasi penting untuk analisis sentimen atau pemrosesan bahasa alami. Dengan menghapus kata-kata ini, analisis dapat lebih fokus pada kata-kata yang memiliki makna kontekstual dan memberikan insight yang lebih mendalam. Penghapusan ini membantu dalam mengurangi kebisingan dalam data teks, sehingga meningkatkan kualitas analisis yang dilakukan.

5. *Stemming*

   Stemming pada penelitian ini menggunakan *library* Sastrawi, yang merupakan Bahasa alami untuk Bahasa Indonesia. Stemming adalah mereduksi kata-kata ke bentuk dasar atau akarnya untuk menganalisis kata-kata yang memiliki makna dasar sama sebagai satu entitas. Misalnya, kata "berjalan", "jalan", dan "berjalannya" akan direduksi menjadi kata dasar "jalan". Ini memudahkan analisis pola dan frekuensi kata dalam dataset teks, serta membantu dalam mengidentifikasi hubungan semantik antar kata.

3.2.4 Data split

Dalam penelitian ini, evaluasi akurasi model sangat penting untuk memastikan keandalan dalam klasifikasi sentimen. Oleh karena itu, penelitian ini mengadopsi teknik *data split* untuk menguji kinerja model menggunakan proporsi yang berbeda dalam pembagian data training dan testing. Metode pemilihan data *random sampling* digunakan untuk memastikan distribusi yang merata dan objektif dari dataset. Dalam rangka menentukan proporsi pembagian data yang paling efektif, penelitian ini mencoba tiga skenario berbeda yaitu 70% data sebagai data training dan 30% sebagai data testing, 90% data *training* dan 10% data *testing*, serta 80% data *training* dan 20% data *testing*.

Hasil pengujian menunjukkan bahwa skenario dengan 80% data *training* dan 20% data *testing* memberikan akurasi terbaik. Keputusan untuk menggunakan pembagian ini didasarkan pada performa optimal yang dicapai dalam skenario ini, dimana model mampu menghasilkan hasil yang paling akurat dalam klasifikasi sentimen. Akurasi yang tinggi pada pembagian ini menegaskan bahwa proporsi 80:20 adalah yang paling sesuai untuk penelitian ini, sehingga akan diadopsi untuk seluruh proses evaluasi model selanjutnya.

3.2.5 TFIDF

Menggunakan teknik TF-IDF, penelitian ini akan menilai pentingnya kata dalam dokumen relatif terhadap seluruh kumpulan data. Teknik ini memberi bobot lebih pada kata-kata yang mungkin menentukan topik atau tema penting dalam teks.

3.2.6 Modeling

Klasifikasi sentimen adalah proses di mana model machine learning digunakan untuk menentukan sentimen dari teks ulasan pengguna. Setelah tahap *Data split*, model akan dilatih menggunakan *training set* dimana setiap ulasan telah diberi label sentimen yang sesuai, seperti positif dan negatif. Dalam penelitian ini, akan digunakan dua algoritma yaitu *Support Vector Machine* (SVM) dan *Naïve Bayes* untuk melakukan klasifikasi sentimen. Model SVM akan mempelajari bagaimana membedakan antara ulasan positif dan negatif dengan mencari *hyperplane* optimal dalam ruang fitur yang memaksimalkan margin antara dua kelas dan mencari akurasi terbaik dari setiap kernelnya. Sedangkan *Naïve Bayes* akan menghitung probabilitas masing-masing kelas berdasarkan fitur yang diberikan dan mengklasifikasikannya sesuai dengan probabilitas terbesar.

3.2.7 *Evaluation*

Setelah model SVM dan *Naïve Bayes* diterapkan pada testing set, kita mengumpulkan hasil prediksi mereka dan membandingkannya dengan label sebenarnya yang telah diberikan sebelumnya pada tahap pelabelan data. Metrik

yang umum digunakan dalam evaluasi model klasifikasi termasuk akurasi, yang mengukur proporsi prediksi yang benar terhadap semua prediksi. Precision menilai seberapa banyak prediksi positif yang sebenarnya benar, recall mengukur seberapa baik model mengidentifikasi semua kasus positif yang sebenarnya, dan *F1-Score* yang merupakan rata-rata harmonis dari precision dan *recall*, memberikan keseimbangan antara keduanya. Selain itu, *Confusion Matrix* akan digunakan untuk memberikan gambaran visual dan detail tentang kinerja model. Confusion Matrix akan menunjukkan jumlah *True Positives* (TP), *True Negatives* (TN), *False Positives* (FP), dan *False Negatives* (FN). Dari data ini, kita dapat mengekstrak insight tentang jenis kesalahan yang cenderung dibuat oleh model dan dalam konteks apa kesalahan tersebut terjadi. Di samping itu, *Area Under the Curve* (AUC) curve juga akan dihitung sebagai salah satu metrik penting untuk menilai kemampuan model dalam membedakan antara kelas positif dan negatif di semua ambang klasifikasi yang mungkin.

3.3 Teknik Pengumpulan Data

Pengumpulan data dalam penelitian ini dilaksanakan dengan memanfaatkan bahasa pemrograman *Python*, yang dijalankan melalui aplikasi *Jupyter Notebook* dan navigator Anaconda. Proses ini bertujuan untuk mengambil data ulasan pengguna aplikasi TikTok yang terdapat pada Google Play Store, sebuah metode yang dikenal sebagai data scraping. Untuk tujuan pengumpulan data ini, digunakan fitur google-play-scraper yang dapat diakses melalui kode *Python* pada *Jupyter Notebook*.

Dengan implementasi kode *Python* tersebut, berhasil dikumpulkan total ulasan sebanyak 1430 data. Seleksi data kemudian dilakukan untuk menyesuaikan dengan periode waktu penelitian yang ditetapkan selama 7 bulan, mulai dari 1 September 2023 hingga 1 April 2024. Hasil seleksi tersebut menghasilkan 1430 data ulasan yang relevan dan sesuai dengan kriteria waktu yang diinginkan.

3.4 Teknik Analisis Data

Pada penelitian ini akan melakukan analisis data yang mengandalkan pendekatan kualitatif untuk menilai secara mendalam berbagai ulasan pengguna

tentang aplikasi TikTok. Pemrosesan dan analisis data ini dilaksanakan dengan memanfaatkan bahasa pemrograman *Python* bersama dengan metode-metode klasifikasi dalam pembelajaran mesin. Penelitian ini menerapkan algoritma *Support Vector Machine* (SVM) dan membandingkannya dengan *Naïve Bayes* dalam rangka mendapatkan hasil yang lebih akurat dan informatif.