

BAB III METODOLOGI PENELITIAN

3.1 Subjek dan Objek Penelitian

Subjek penelitian adalah individu, objek, atau lokasi yang menjadi fokus dari penelitian atau yang menjadi topik utama dalam penelitian [44]. Dalam penelitian ini, subjek penelitiannya adalah pengguna aplikasi Vidio yang menulis ulasan di *Google Play Store*.

Objek penelitian adalah materi, entitas, atau hal lain yang menjadi pusat perhatian dan penelitian [44]. Dalam penelitian ini, objek penelitiannya adalah sentimen pengguna pada ulasan aplikasi Vidio dan aspek-aspek yang memengaruhi sentimen tersebut.

3.2 Alat dan Bahan Penelitian

3.2.1 Alat Penelitian

Untuk menunjang penelitian, alat penelitian yang digunakan berupa *hardware* dan *software* sebagai berikut:

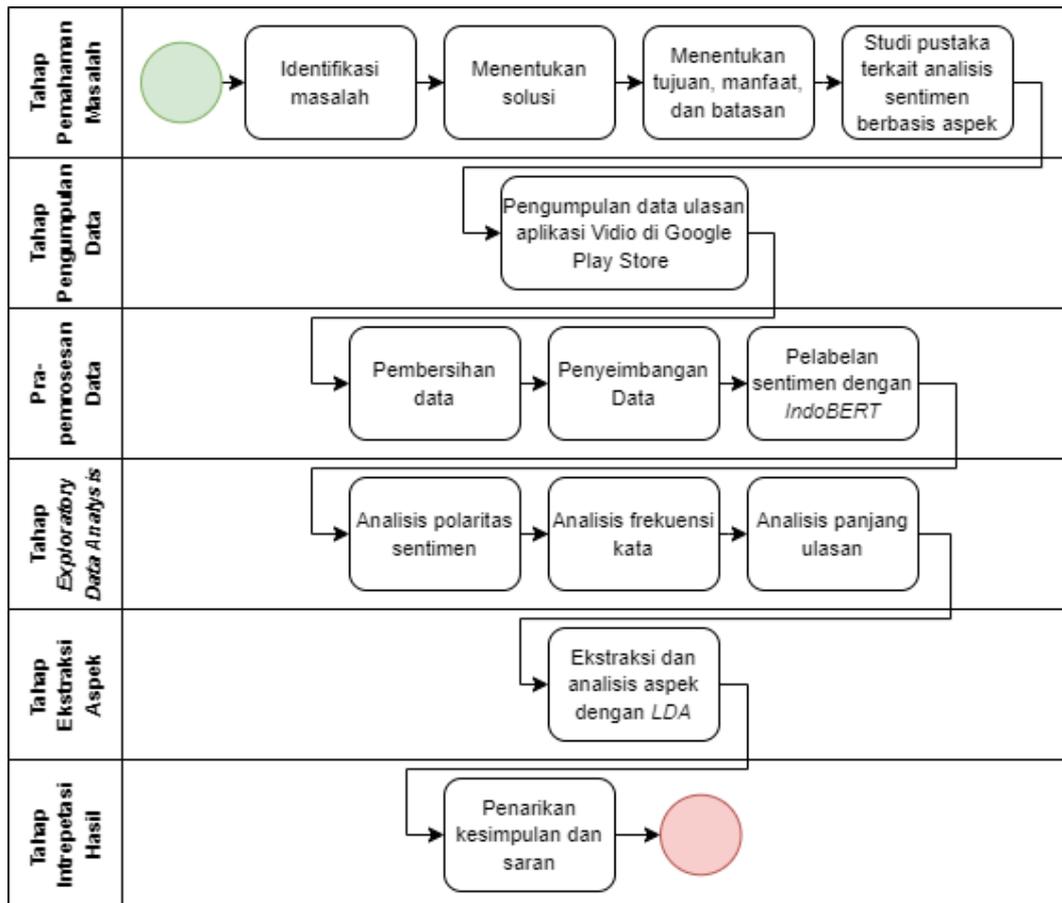
- 1) *Hardware* (perangkat keras)
 - a) Laptop Acer Swift X SFX14
 - b) Mouse
- 2) *Software* (perangkat lunak)
 - a) Windows 10 Home Edition
 - b) Google Chrome
 - c) Jupyter Notebook
 - d) Python 3.9.1
 - e) Microsoft Office 2019
 - f) Mendeley

3.2.2 Bahan Penelitian

Bahan penelitian yang digunakan adalah data hasil *web scraping* aplikasi Vidio dari *Google Play Store*. Data ini berisi daftar dan informasi ulasan pengguna dari tahun 2014 sampai tahun 2023. Kolom yang digunakan adalah fitur ulasan.

3.3 Diagram Alir Penelitian

Diagram alir penelitian memainkan peran penting dalam memastikan bahwa penelitian dilakukan dengan urut dan tidak mengganggu tahapan penelitian lainnya. Ilustrasi tahapan-tahapan penelitian yang akan ditempuh diilustrasikan seperti Gambar 3.1.



Gambar 3.1 Diagram Alir Penelitian

3.3.1 Tahap Pemahaman Masalah

Tahap pemahaman masalah penelitian ini terdiri atas empat bagian, yaitu identifikasi masalah, penentuan solusi, penentuan tujuan, manfaat, dan batasan masalah, dan studi pustaka terkait analisis sentimen berbasis aspek. Penelitian ini dimulai dengan mengidentifikasi masalah utama yang berkaitan dengan feedback pengguna dalam bentuk ulasan.

Melalui observasi terhadap ulasan aplikasi Vidio di *Google Play Store*, ditemukan bahwa banyak ulasan yang secara umum maupun spesifik mengandung

nuansa kritikan ataupun keluhan. Melalui hal tersebut, penelitian ini bertujuan untuk mengkaji lebih dalam tentang bagaimana sentimen dalam ulasan yang ditulis pengguna tersebut dan aspek-aspek apa saja yang mempengaruhinya, dengan tujuan utama untuk menggali polaritas sentimen ulasannya dan aspek terkait yang mempengaruhi.

Penelitian ini diharapkan dapat memberikan *insight* berharga bagi pengembang aplikasi Vidio untuk meningkatkan kualitas layanan mereka berdasarkan umpan balik pengguna. Untuk mendukung tujuan tersebut, penelitian ini juga membatasi fokusnya pada ulasan aplikasi Vidio yang berbahasa Indonesia dari periode tahun 2014 hingga tahun 2023. Penelitian ini juga membatasi untuk tidak melakukan pengukuran keakuratan hasil pelabelan. Ini karena penelitian berfokus pada pemahaman umum sentimen dan aspek tanpa perlu mengkonfirmasi setiap detail label secara akurat.

Dalam upaya untuk mengatasi masalah yang telah diidentifikasi, penelitian ini mengadopsi penggunaan model *pre-trained IndoBERT* dan *LDA*. *IndoBERT* merupakan model *deep learning* yang dirancang khusus untuk memahami bahasa Indonesia secara efektif dan efisien, memberikan kemampuan untuk menganalisis sentimen dan aspek dalam ulasan dengan lebih akurat.

Selain itu, penelitian ini juga mengkaji studi pustaka yang relevan dan sesuai terkait analisis sentimen berbasis aspek untuk mendukung implementasi teori dan konsep yang terkait dengan aspek-aspek yang memengaruhi pengalaman pengguna aplikasi Vidio. Kajian ini dimaksudkan untuk memperkuat landasan teori dan konsep yang digunakan dalam penelitian, termasuk pemahaman tentang bagaimana pengalaman pengguna bisa dipengaruhi oleh berbagai aspek berdasarkan analisis ulasannya.

3.3.2 Tahap Pengumpulan Data

Dalam tahapan penelitian ini, dilakukan pengumpulan data dari ulasan aplikasi Vidio di *Google Play Store* selama periode tahun 2014 sampai tahun 2023. Data tersebut diambil dengan menggunakan teknik *web scraping*, yaitu teknik pengambilan data dari situs web melalui program komputer. Untuk melakukan *web scraping*, penelitian ini menggunakan bahasa pemrograman Python dan modul

google-play-scraper yang dapat mengekstraksi informasi dari aplikasi Vidio di *Google Play Store*. Data ulasan yang diperoleh melalui cara tersebut akan terdiri dari 11 variabel, yaitu 9 variabel kategorikal dan 2 variabel numerik. Variabel-variabel ini dapat dilihat pada Tabel 3.1.

Tabel 3.1 Spesifikasi Variabel Dataset Penelitian

Variabel	Deskripsi	Tipe Data
<i>reviewId</i>	Kode unik yang mengidentifikasi setiap ulasan.	Kategorikal
<i>userName</i>	Nama pengguna yang memberikan ulasan.	Kategorikal
<i>userImage</i>	Gambar profil pengguna yang memberikan ulasan.	Kategorikal
<i>content</i>	Isi teks dari ulasan.	Kategorikal
<i>score</i>	Nilai rating yang diberikan oleh pengguna, dari 1 sampai 5.	Numerikal
<i>thumbsUpCount</i>	Jumlah pengguna lain yang menyukai ulasan tersebut.	Numerikal
<i>reviewCreatedVersion</i>	Versi aplikasi Vidio yang digunakan pengguna saat memberikan ulasan.	Kategorikal
<i>at</i>	Tanggal dan waktu ulasan dibuat.	Kategorikal
<i>replyContent</i>	Isi teks dari balasan yang diberikan oleh pengembang aplikasi Vidio.	Kategorikal
<i>repliedAt</i>	Tanggal dan waktu balasan dibuat.	Kategorikal
<i>appVersion</i>	Versi aplikasi Vidio yang terbaru saat ulasan dibuat.	Kategorikal

3.3.3 Pra-pemrosesan Data

Tahapan ini bertujuan untuk mempersiapkan data teks sebelum melakukan analisis lanjutan, seperti analisis sentimen dan ekstraksi aspek. Tahap ini terdiri dari beberapa sub-tahapan, yaitu pembersihan data, penyeimbangan data, dan pelabelan sentimen dengan *IndoBERT*.

Tahap ini juga bertujuan untuk menyesuaikan data teks dengan karakteristik atau tujuan penelitian dan analisis yang akan dilakukan. Data teks yang sebelumnya dikumpulkan berupa ulasan pengguna aplikasi Vidio di *Google Play Store*. Dari data

tersebut, hanya fitur ulasan (variabel 'content') yang akan digunakan, sedangkan fitur lain akan diabaikan atau dilakukan *drop column*. Proses ini diilustrasikan pada Gambar 3.2.

Data Asli

reviewId	userName	userImage	content	score	thumbsUpCount
reviewCreatedVersion	at	replyContent	replyAt	appVersion	

Seleksi Fitur

reviewId	userName	userImage	content	score	thumbsUpCount
reviewCreatedVersion	at	replyContent	replyAt	appVersion	

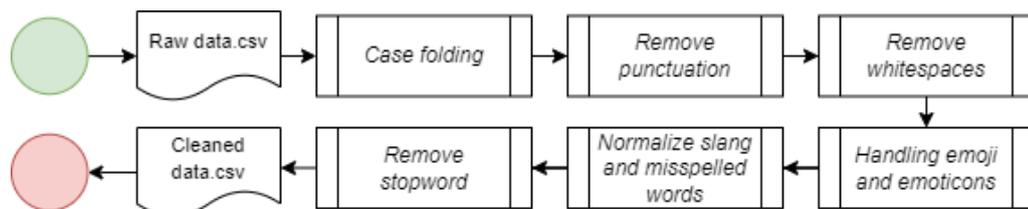
Data Final

content

Gambar 3.2 Seleksi Fitur

3.3.3.1 Pembersihan Data

Sub-tahapan ini bertujuan untuk membersihkan data teks dari *noise* (gangguan) yang dapat mengurangi kualitas dari analisis sentimen berbasis aspek. *Noise* dapat berupa simbol, tanda baca, angka, spasi berlebih, dan kata-kata yang tidak relevan. Selain itu, sub-tahapan ini juga bertujuan untuk memperbaiki kata-kata yang tidak baku, salah eja, atau mengandung *typo*, dengan menggunakan kamus atau aturan bahasa yang sesuai. Untuk melakukan pembersihan data teks, akan menggunakan pustaka seperti *nlTK*, *sastrawi*, atau pustaka lain yang sesuai. Alur pembersihan data tersebut diilustrasikan seperti Gambar 3.2.



Gambar 3.3 Alur Pembersihan Data

1) *Case Folding*.

Proses ini akan mengubah semua huruf dalam data teks menjadi huruf kecil, sehingga tidak ada perbedaan interpretasi akibat jenis ukuran huruf. Contoh penerapan proses ini adalah seperti pada Tabel 3.2.

Tabel 3.2 Contoh *Case Folding*

Sebelum	Sesudah
“Aplikasinya bagus banget, bisa nonton FILM dan BOLA gratis. “	“aplikasinya bagus banget, bisa nonton film dan bola gratis. “
“Bagian ini digunakan untuk mengecilkan ukuran huruf. “	“bagian ini digunakan untuk mengecilkan ukuran huruf. “

2) *Removing punctuation*.

Proses ini akan menghapus tanda baca dalam teks, seperti koma, titik, tanda tanya, dan tanda lain yang tidak memberikan informasi penting untuk analisis sentimen berbasis aspek. Contoh penerapan proses ini adalah seperti pada Tabel 3.3.

Tabel 3.3 Contoh *Removing Punctuation*

Sebelum	Sesudah
“aplikasinya bagus banget, bisa nonton film dan bola gratis. “	“aplikasinya bagus banget bisa nonton film dan bola gratis “
“bagian ini digunakan untuk mengecilkan ukuran huruf. “	“bagian ini digunakan untuk mengecilkan ukuran huruf “

3) *Removing whitespace*.

Proses ini akan menghapus *whitespace* (spasi berlebih) dalam teks, agar teks menjadi lebih rapi dan mudah dibaca. Hal ini dapat membantu menghindari kesalahan dalam pemrosesan teks selanjutnya. Contoh penerapan proses ini adalah seperti pada Tabel 3.4.

Tabel 3.4 Contoh *Removing Whitespace*

Sebelum	Sesudah
“aplikasinya bagus banget nonton film bola gratis “	“aplikasinya bagus banget nonton film bola gratis”
“bagian digunakan mengecilkan ukuran huruf “	“bagian digunakan mengecilkan ukuran huruf”

4) *Handling emoji and emoticons.*

Proses ini bertujuan untuk mengubah emoji dan emotikon menggunakan *regex* dan nilai *unicode* dari emoji. Contoh penerapan proses ini adalah seperti pada Tabel 3.5.

Tabel 3.5 Contoh *Handling Emoji and Emoticon*

Sebelum	Sesudah
“aplikasinya bagus banget nonton film bola gratis :)”	“aplikasinya bagus banget nonton film bola gratis”
“bagian digunakan mengecilkan ukuran huruf 😊”	“bagian digunakan mengecilkan ukuran huruf”

5) *Normalizing slang and misspelled words.*

Proses ini bertujuan untuk mengubah kata-kata yang tidak baku, salah eja, atau mengandung *typo* menjadi kata-kata yang baku dan benar. Proses ini akan melibatkan penggunaan kamus slang yang diambil dari GitHub. Kamus slang ini kemudian diubah sesuai dengan KBBI (Kamus Besar Bahasa Indonesia) untuk menormalisasikan kata-kata singkat dan tidak baku menjadi kata baku. Contoh penerapan proses ini adalah seperti pada Tabel 3.6.

Tabel 3.6 Contoh *Normalizing Slang and Misspelled*

Sebelum	Sesudah
“aplikasinya bgs bgt nonton film bola gratis”	“aplikasinya bagus sekali nonton film bola gratis”
“bagian digunakan mengecilkan ukrn huruf”	“bagian digunakan mengecilkan ukuran huruf”

6) *Removing stopword.*

Proses ini akan menghapus kata-kata umum dalam teks, seperti kata sambung, kata depan, kata ganti, dan kata lain yang tidak memberikan makna. Contoh penerapan proses ini adalah seperti pada Tabel 3.7.

Tabel 3.7 Contoh *Removing Stopword*

Sebelum	Sesudah
“aplikasinya bagus banget bisa nonton film dan bola gratis “	“aplikasinya bagus banget nonton film bola gratis “

“bagian ini digunakan untuk mengecilkan ukuran huruf “	“bagian digunakan mengecilkan ukuran huruf “
--	--

3.3.3.2 Penyeimbangan Data

Sub-tahapan ini bertujuan untuk menyesuaikan distribusi data agar dataset menjadi seimbang berdasarkan jumlah data per tahun. Hal ini penting untuk memastikan bahwa analisis yang dilakukan tidak bias akibat distribusi data yang tidak merata [45]. Dalam konteks ini, penyeimbangan data dilakukan dengan menggunakan pendekatan *undersampling*.

Undersampling adalah teknik penyeimbangan data yang mengurangi jumlah sampel dalam kelas mayoritas untuk menyamakan jumlahnya dengan kelas minoritas. Pendekatan ini dipilih karena dengan mengatur semua tahun pada jumlah data minimal yang konsisten, akan menciptakan kondisi yang sama untuk setiap tahun. Hal ini secara signifikan mengurangi potensi bias pada analisis selanjutnya yang disebabkan oleh perbedaan volume data [46][47].

Proses *undersampling* dilakukan dengan langkah-langkah berikut:

- 1) Identifikasi jumlah data per tahun: Menghitung jumlah data yang tersedia untuk setiap tahun. Ini dilakukan dengan mengelompokkan data berdasarkan tahun dan menghitung jumlah entri untuk setiap kelompok.
- 2) Menentukan tahun dengan jumlah data minimal: Menentukan tahun dengan jumlah data paling sedikit. Tahun ini akan menjadi acuan untuk jumlah data yang diambil dari tahun-tahun lainnya.
- 3) Pengurangan sampel untuk tahun lainnya: Mengurangi jumlah sampel dari tahun-tahun lainnya hingga jumlahnya sama dengan jumlah data dari tahun dengan jumlah data minimal. Pengurangan ini dilakukan secara acak untuk memastikan representasi data tetap terjaga.

3.3.3.3 Pelabelan Sentimen dengan *IndoBERT*

Sub-tahapan ini bertujuan untuk memberikan label sentimen pada setiap ulasan dalam data teks menggunakan model *pre-trained IndoBERT*. Model ini menerima input berupa teks ulasan yang disesuaikan dengan representasi input *IndoBERT*. Proses penyesuaian input secara otomatis meliputi langkah-langkah berikut:

- 1) Mengubah hasil *preprocessing* menjadi token-token kata. Contoh penerapan

proses ini adalah seperti pada Tabel 3.8.

Tabel 3.8 Contoh Perubahan menjadi Token *IndoBERT*

Sebelum	Sesudah
“aplikasinya bagus sekali nonton film bola gratis”	[‘aplikasi’, ‘bagus’, ‘sekali’, ‘nonton’, ‘film’, ‘bola’, ‘gratis’]
“bagian digunakan mengecilkan ukuran huruf”	[‘bagian’, ‘digunakan’, ‘mengecilkan’, ‘ukuran’, ‘huruf’]

- 2) Menambahkan token [CLS] dan [SEP] sebagai pemisah antar kalimat. Contoh penerapan proses ini adalah seperti pada Tabel 3.9.

Tabel 3.9 Contoh Penambahan Token [CLS] dan [SEP]

Sebelum	Sesudah
[‘aplikasi’, ‘bagus’, ‘sekali’, ‘nonton’, ‘film’, ‘bola’, ‘gratis’]	[‘[CLS]’, ‘aplikasi’, ‘bagus’, ‘sekali’, ‘nonton’, ‘film’, ‘bola’, ‘gratis’, ‘[SEP]’]
[‘bagian’, ‘digunakan’, ‘mengecilkan’, ‘ukuran’, ‘huruf’]	[‘[CLS]’, ‘bagian’, ‘digunakan’, ‘mengecilkan’, ‘ukuran’, ‘huruf’, ‘[SEP]’]

- 3) Penyesuaian panjang kalimat dengan panjang maksimal input *BERT*, yaitu 512 token. Contoh penerapan proses ini adalah seperti pada Tabel 3.10.

Tabel 3.10 Contoh Penyesuaian Panjang Kalimat

Sebelum	Sesudah
[‘[CLS]’, ‘aplikasi’, ‘bagus’, ‘sekali’, ‘nonton’, ‘film’, ‘bola’, ‘gratis’, ‘[SEP]’]	[‘[CLS]’, ‘aplikasi’, ‘bagus’, ‘banget’, ‘nonton’, ‘film’, ‘bola’, ‘gratis’, ‘[SEP]’, ‘[PAD]’, ‘[PAD]’, ..., ‘[PAD]’]
[‘[CLS]’, ‘bagian’, ‘digunakan’, ‘mengecilkan’, ‘ukuran’, ‘huruf’, ‘[SEP]’]	[‘[CLS]’, ‘bagian’, ‘digunakan’, ‘mengecilkan’, ‘ukuran’, ‘huruf’, ‘[SEP]’, ‘[PAD]’, ‘[PAD]’, ..., ‘[PAD]’]

- 4) Memisahkan *subword* (sub kata) pada kata yang tidak ada di *vocabulary* (*out-of-vocabulary*). Misalnya, jika kata ‘gratis’ tidak ada dalam *vocabulary*, maka akan dipecah menjadi sub kata ‘gra’, ‘##tis’. Contoh penerapan proses ini adalah seperti pada Tabel 3.11.

Tabel 3.11 Contoh Pemisahan *Subword*

Sebelum	Sesudah
[‘[CLS]’, ‘aplikasi’, ‘bagus’,	[‘[CLS]’, ‘aplikasi’, ‘bagus’,

'banget', 'nonton', 'film', 'bola', 'gratis', '[SEP]'	'banget', 'nonton', 'film', 'bola', 'gra', '##tis', '[SEP]', '[PAD]', ..., '[PAD]'
['[CLS]', 'bagian', 'digunakan', 'mengecilkan', 'ukuran', 'huruf', '[SEP]', '[PAD]', '[PAD]', ..., '[PAD]']	['[CLS]', 'bagian', 'digunakan', 'mengecilkan', 'ukuran', 'huruf', '[SEP]', '[PAD]', '[PAD]', ..., '[PAD]'] (tidak ada perubahan karena semua kata ada dalam <i>vocabulary</i>)

- 5) Menentukan *attention mask* untuk memisahkan nilai token kata dan nilai *padding*. Token *padding* [PAD] menjadi nilai 0, sedangkan token kata menjadi nilai 1. Penerapan proses ini adalah seperti pada Tabel 3.12.

Tabel 3.12 Contoh Penentuan *Attention Mask*

Sebelum	Sesudah
['[CLS]', 'aplikasi', 'bagus', 'banget', 'nonton', 'film', 'bola', 'gra', '##tis', '[SEP]', '[PAD]', ..., '[PAD]']	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, ..., 0]
['[CLS]', 'bagian', 'digunakan', 'mengecilkan', 'ukuran', 'huruf', '[SEP]', '[PAD]', '[PAD]', ..., '[PAD]']	[1, 1, 1, 1, 1, 1, 1, 0, 0, ..., 0]

- 6) Menyiapkan data *loader* untuk setiap data, yaitu data *loader* untuk pelatihan, pengujian, dan validasi.
- 7) Setelah input disesuaikan, model *pre-trained IndoBERT* bisa digunakan untuk memproses token-token tersebut dengan menggunakan *encoder* dan *decoder* dari *layer transformer*.

3.3.4 Tahap *Exploratory Data Analysis (EDA)*

Tahap *EDA* merupakan bagian yang krusial dalam penelitian ini. Tujuan utamanya adalah untuk menyelidiki pola dan hubungan yang ada dalam data, serta untuk menguji asumsi-asumsi yang menjadi fondasi analisis. Meskipun dalam praktik umum, *EDA* sering dilakukan sebelum pembersihan data, dalam penelitian ini, *EDA* justru dilakukan setelah pembersihan data. Hal ini dilakukan untuk memastikan kualitas data yang akan dianalisis. Dengan demikian, tahap *EDA* diharapkan dapat berfokus pada elemen data yang benar-benar relevan dan

memberikan hasil analisis yang lebih akurat dan mendalam. Pada penelitian ini, tahap *EDA* terdiri dari tiga sub-tahap, yaitu analisis polaritas sentimen, analisis frekuensi kata, dan analisis panjang ulasan.

3.3.4.1 Analisis Polaritas Sentimen

Sub-tahap ini dilakukan untuk mengevaluasi distribusi dan perbandingan antara sentimen positif, negatif, dan netral dalam data pada setiap tahunnya. Sub-tahap ini dapat memberikan gambaran tentang tingkat kepuasan dan persepsi pengguna terhadap aplikasi pada setiap tahunnya. Variabel sentimen yang telah dibuat pada tahap pelabelan, kemudian dianalisis secara keseluruhan dengan menghitung jumlah dari masing-masing sentimen dan menampilkan visualisasi seperti *bar chart* untuk menunjukkan proporsi sentimen per tahun dan *pie chart* untuk proporsi secara keseluruhan. Hasil analisis polaritas secara keseluruhan dalam bentuk rekapan informasi juga dapat diilustrasikan seperti Tabel 3.13.

Tabel 3.13 Perbandingan Polaritas Opini Pengguna Aplikasi Vidio

Polaritas Opini	Jumlah Ulasan	Persentase
Negatif
Positif
Netral
Total

3.3.4.2 Analisis Frekuensi Kata

Sub-tahap ini bertujuan untuk menghitung dan memberikan gambaran jumlah kemunculan kata dalam ulasan. Frekuensi kata memberikan informasi tentang kepentingan utama pengguna terhadap aplikasi. Analisis ini juga dapat digunakan untuk mengidentifikasi dan menghapus kata-kata yang tidak relevan untuk meningkatkan kualitas dan efisiensi data saat melakukan ekstraksi aspek.

Untuk melakukan analisis frekuensi kata, sub-tahap ini memanfaatkan library '*collections*' yang tersedia di Python. Sub-tahapan ini terdiri atas dua analisis berikut:

- 1) Analisis frekuensi kata per sentimen, yang bertujuan untuk membandingkan kata-kata yang sering muncul di setiap kategori sentimen.

- 2) Analisis frekuensi kata keseluruhan, yang bertujuan untuk mengidentifikasi kata-kata yang sering muncul secara keseluruhan.

3.3.4.3 Analisis Panjang Ulasan

Sub-tahap ini bertujuan untuk mengukur dan membandingkan jumlah kata dalam sebuah ulasan. Hal ini dapat memberikan gambaran tentang seberapa besar minat atau motivasi pengguna untuk berbagi pengalaman atau pendapatnya tentang aplikasi lewat panjang atau pendeknya suatu ulasan. Sub-tahap ini juga dapat membantu menilai tingkat kedalaman ulasan.

Sub-tahap ini menggunakan histogram untuk menampilkan distribusi panjang ulasan dalam jumlah karakter. Histogram dapat menunjukkan sebaran ulasan yang pendek, sedang, atau panjang. Sub-tahap ini juga dapat menganalisis ulasan yang paling informatif dan relevan untuk diekstraksi aspeknya. Ulasan yang panjang biasanya mengandung lebih banyak informasi dan pendapat daripada ulasan yang sederhana atau singkat.

3.3.5 Tahap Ekstraksi Aspek

Tahapan ekstraksi aspek bertujuan untuk mengidentifikasi dan mengelompokkan aspek-aspek yang menjadi fokus perhatian atau sumber sentimen dari pengguna aplikasi Vidio di *Google Play Store*. Untuk melakukan tahap ini, data teks harus diubah menjadi bentuk vektor, yaitu representasi numerik dari teks yang dapat diproses oleh algoritma. Dalam tahap ini, LDA secara otomatis mengubah data teks menjadi bentuk vektor, yaitu representasi numerik dari teks yang dapat diproses oleh algoritma. Proses ini mengintegrasikan segmentasi, vektorisasi, dan ekstraksi aspek dalam satu langkah, sehingga tidak memerlukan sub-tahapan tambahan.

3.3.5.1 Segmentasi

Sub-tahapan ini bertujuan untuk membagi data teks menjadi unit yang lebih kecil, yaitu kalimat dan kata, agar dapat dilakukan analisis sentimen berbasis aspek. Namun, dalam penelitian ini, data teks akan langsung disegmentasi menjadi kata atau token dengan menggunakan spasi sebagai pemisah. Proses tokenisasi akan dilakukan menggunakan pustaka *nltk*, yang menyediakan fungsi-fungsi untuk melakukan segmentasi teks dengan mudah dan cepat.

3.3.5.2 Vektorisasi

Sub-tahapan ini bertujuan untuk mengubah data teks menjadi vektor yang terdiri dari pasangan kata dan frekuensinya. Pada penelitian ini, vektorisasi dilakukan lewat pustaka *gensim* dengan metode *TF-IDF* dari *scikit-learn*.

3.3.5.3 Ekstraksi dan analisis aspek dengan LDA

Tahapan ini bertujuan untuk mengidentifikasi dan mengelompokkan aspek-aspek yang menjadi fokus perhatian atau sumber sentimen dari pengguna menggunakan model *LDA* (*Latent Dirichlet Allocation*). Langkah-langkah penerapan *LDA* untuk analisis berbasis aspek pada penelitian ini adalah sebagai berikut:

- 1) Memuat model dan korpus dari repositori *indolem* atau *indobenchmark* sebagai input untuk *LDA* dengan menginstal pustaka *gensim*.
- 2) Membuat *dictionary* dan korpus dengan mengubah data menjadi bentuk vektor. Lalu mencari nilai koherensi menggunakan fungsi '*CoherenceModel*'.
- 3) Membangun model *LDA* dengan memasukkan jumlah topik sesuai dengan nilai koherensi yang optimal menggunakan fungsi *LdaMulticore*.
- 4) Mencetak kata-kata kunci beserta bobotnya menggunakan fungsi '*print_topics*'. Lalu membuat matriks kesamaan berdasarkan *word embeddings* yang dihasilkan oleh model *FastText* dari *gensim*.
- 5) Membuat interpretasi visual hasil analisis kluster topik menggunakan pustaka *PyLDAvis*.

Hasil dari tahapan ini adalah kelompok-kelompok kata atau kluster yang merepresentasikan topik-topik dalam teks ulasan. Setiap kluster kemudian dianalisis secara manual untuk mengidentifikasi kata-kata atau topik yang terkandung di dalamnya dan menghubungkannya dengan sentimen lewat visualisasi 30 kata teratas dalam setiap topik dari pustaka *PyLDAvis*.

3.3.6 Tahap Interpretasi Hasil

Tahap ini bertujuan untuk menampilkan hasil-hasil yang didapatkan dari penelitian. Tahap ini juga mencakup proses menyimpulkan dan memberikan saran penelitian. Kesimpulan dan saran dibuat berdasarkan hasil interpretasi yang telah

diuraikan pada tahapan sebelumnya. Tahap ini juga menampilkan ringkasan dari keseluruhan penelitian yang telah dilaksanakan, termasuk latar belakang, tujuan, metode, dan hasil penelitian. Ringkasan ini juga akan menjawab pertanyaan penelitian yang telah dirumuskan berikut:

- 1) Perbandingan polaritas opini pengguna aplikasi Vidio.
- 2) Identifikasi aspek-aspek yang berpengaruh terhadap opini pengguna aplikasi Vidio.

Dengan menyelesaikan semua kegiatan penelitian sesuai dengan tahapan yang sudah ditentukan, maka penelitian ini dianggap selesai dilaksanakan.