`

# BAB II
# TINJAUAN PUSTAKA

## 2.1 Kajian Pustaka

Penelitian ini bertujuan untuk melakukan analisis sentimen pengguna dan aspek-aspek yang memengaruhi sentimen tersebut pada aplikasi Vidio yang tersedia di *Google Play Store* dengan menggunakan model *pre-trained*. Sebelum melakukan analisis tersebut, perlu melengkapi data dan mempertajam rumusan masalah penelitian dengan mengkaji literatur-literatur yang relevan dengan topik dan tema penelitian ini. Kajian literatur ini meliputi beberapa jurnal yang dipilih berdasarkan kriteria kesesuaian dan keterkaitan dengan penelitian ini.

Dari hasil kajian literatur, ditemukan bahwa penelitian tentang analisis sentimen berbasis aspek pada ulasan aplikasi Vidio di *Google Play Store* masih jarang dilakukan. Namun beberapa penelitian terdahulu terkait analisis sentimen berbasis aspek adalah seperti penelitian [12], yang menggunakan *LDA, Semantic Similarity*, dan *BERT* untuk menganalisis sentimen pengguna hotel berdasarkan 5 aspek yaitu '*Comfort*', '*Cleanliness'*, '*Location*', '*Service*', dan '*Food*'. Aspek-aspek tersebut dibuat berdasarkan pola kata kunci dari penelitian lain terkait perhotelan dan *BERT*.

Berdasarkan penelitian [12], ekstraksi aspek dengan *BERT* dan *Semantic Similarity* dapat menghasilkan model anotasi sentimen dengan kinerja presisi 86%, *recall* 92%, dan *f1-score* 89%. Sementara itu, analisis sentimen dengan *BERT* dapat menghasilkan kinerja presisi 96%, *recall* 98%, dan *f1-score* 97%. Walaupun *BERT* dan *Semantic Similarity* memiliki performa yang baik dalam melakukan ekstraksi aspek dan analisis sentimen, tetapi *BERT* memerlukan waktu pelatihan yang cukup lama. Oleh karena itu, penelitian [12], juga menyarankan untuk menggunakan *GPU* yang cukup memadai agar efisien untuk menjalankan *BERT*.

Penelitian lain yang terkait dengan topik ini adalah penelitian [13], yang mengusulkan model *Convolutional Neural Network* (*CNN*) untuk menganalisis sentimen berbasis aspek pada ulasan aplikasi PeduliLindungi yang tersedia di *Google Play Store*. Dalam penelitian ini, penulis melakukan pelabelan aspek dengan

mengikuti standar kategori aspek dari Android, yang meliputi *Visual Experience* (*UI*/*UX*), *Functionality*, *Performance and Stability*, dan *Privacy and Security*. Selain itu, penelitian ini juga membuat aspek *Functionality* menjadi fitur di aplikasi PeduliLindungi.

Hasil evaluasi dari model penelitian [13] menunjukkan kinerja *f1-score* 92.23% dalam klasifikasi aspek dan 95.13% dalam klasifikasi sentimen. Dari hasil analisis sentimen, terlihat bahwa mayoritas pengguna memiliki sentimen negatif terhadap aspek '*Visual Experience*', '*Scan-Checkin/Out*', '*Vaccine Certificate*', '*eHac*', '*COVID-19 Test*', '*Register/Login*', '*Performance and Stability*', and '*Privacy*, *Data*, and *Security*', yang mengindikasikan bahwa aplikasi PeduliLindungi masih perlu banyak perbaikan. Namun, penelitian ini memiliki keterbatasan dalam hal keseimbangan data, dan tidak menjelaskan bagaimana mengatasi masalah tersebut.

Penelitian terkait analisis sentimen berbasis aspek lainnya adalah penelitian [14]. Penelitian ini menggunakan data ulasan Hotel Tentrem Yogyakarta, yang merupakan hotel populer bagi para wisatawan yang berkunjung ke Yogyakarta, sebagai objek penelitian. Data ulasan tersebut diperoleh dari situs TripAdvisor, yang merupakan situs web multilingual. Penelitian ini mengklasifikasikan polaritas sentimen pengguna terhadap lima aspek utama yang berpengaruh terhadap kualitas hotel, yaitu pelayanan, lokasi, kamar, kolam renang, dan pusat kebugaran. Aspek-aspek tersebut ditentukan sebelumnya berdasarkan data ulasan yang ada.

Penelitian [14] menggunakan algoritma *Random Forest Classifier* untuk mengklasifikasikan sentimen berbasis aspeknya dan melakukan pembobotan kata *Term Frequency-Inverse Document Frequency* (*TF-IDF*) untuk mengubah data teks menjadi vektor. Data yang digunakan adalah data ulasan pada Hotel Tentrem di Yogyakarta berbahasa Indonesia yang diperoleh melalui situs TripAdvisor.

Penelitian [14] menerapkan pengujian dengan berbagai skenario parameter kedalaman *tree* dan jumlah *tree*. Hasil pengujian menunjukkan bahwa semakin besar kedalaman *tree* dan jumlah *tree*, maka hasil prediksi akan semakin baik. Hasil klasifikasi terbaik untuk kedua parameter terhadap aspek kamar adalah prediksi dengan kinerja 90% untuk nilai akurasi dan skor f1-nya. Namun, penelitian ini

memiliki beberapa keterbatasan, yaitu keseimbangan data yang rendah, tidak adanya landasan pemilihan lima aspek tersebut untuk klasifikasinya, dan tidak adanya analisis lebih lanjut untuk empat aspek lainnya.

Penelitian terkait selanjutnya adalah penelitian [15], yang mengusulkan delapan model *pre-trained deep learning* yang berbeda dalam analisis sentimen berbasis aspek pada ulasan hotel Indonesia. Penelitian ini memiliki dua skenario klasifikasi sentimen, yaitu klasifikasi sentimen pada semua aspek dan klasifikasi sentimen pada masing-masing aspek. Penelitian ini menggunakan data ulasan hotel Indonesia yang dilabeli secara manual dan diverifikasi oleh seorang ahli yang telah bekerja di sektor perhotelan selama dua tahun.

Penelitian [15] menemukan bahwa model *Long Short-Term Memory* (*LSTM*) adalah model terbaik untuk klasifikasi aspek, sedangkan model *Convolutional Neural Networks* (*CNN*) adalah model terbaik untuk klasifikasi sentimen. Selain itu, penelitian ini juga menganalisis sentimen pengguna terhadap aspek 'hotel, seperti '*price*', 'hotel', '*room*', '*location*', '*service*', dan '*restaurant'*. Hasilnya menunjukkan bahwa sebagian besar pengguna memberikan sentimen positif terhadap aspek-aspek tersebut.

Namun, penelitian [15] juga memiliki beberapa keterbatasan, antara lain keseimbangan data yang rendah, yang mengakibatkan banyak kesalahan klasifikasi, terutama pada skenario klasifikasi sentimen pada semua aspek. Hal ini disebabkan oleh adanya beberapa kata yang memiliki label sentimen yang berbeda dalam satu dokumen. Selain itu, penelitian ini juga memerlukan spesifikasi perangkat yang tinggi untuk menjalankan delapan *pre-trained deep learning* yang berbeda, sehingga kurang efisien dan praktis.

Penelitian terkait analisis sentimen berbasis aspek lainnya adalah penelitian [16], yang menganalisis sentimen pengguna terhadap produk kecantikan di web Female Daily, yang menggunakan bahasa multilingual. Untuk mengklasifikasikan sentimen pengguna, penelitian ini memanfaatkan pembobotan *TF-IDF* dan algoritma *Naïve Bayes*. Empat aspek yang menjadi fokus penelitian ini adalah 'Harga', 'Kemasan', 'Produk', dan 'Aroma'. Keempat aspek tersebut diberi label secara manual oleh empat orang yang berbeda.

Penelitian [16] mencoba tiga skenario berbeda untuk menangani data multilingual. Skenario pertama adalah menggunakan data asli tanpa terjemahan. Skenario kedua adalah menerjemahkan data ke bahasa Inggris. Skenario ketiga adalah menerjemahkan data ke bahasa Inggris, kemudian kembali ke bahasa asal. Hasil penelitian menunjukkan bahwa skenario ketiga memberikan performa terbaik dengan *f1-score* sebesar 62,81%.

Penelitian terkait sentimen berbasis aspek lainnya adalah penelitian [8], yang menguji pengaruh atribut atau aspek restoran terhadap kepuasan pelanggan pada berbagai tipe restoran. Penelitian ini menggunakan model *Valence Aware Dictionary and sEntiment Reasoner* (*VADER*) untuk mengklasifikasikan sentimen dari ulasan pelanggan restoran di Google Maps. Model *VADER* dapat menentukan aspek yang dibahas dalam ulasan secara otomatis dengan menggunakan metode *lexicon-based* (berbasis leksikon) atau *rule-based* (berbasis aturan) dari daftar kata-kata yang disusun oleh peneliti.

Aspek yang dianalisis dalam penelitian [8] ada empat, yaitu '*Food'* (makanan), '*Services'* (pelayanan), '*Atmosphere'* (suasana), dan '*Value*' (harga). Penulis membuat daftar kata-kata untuk setiap aspek agar dapat mengetahui aspek mana yang disebutkan dalam setiap kalimat dari setiap ulasan. penelitian ini memilih kata-kata yang sering muncul untuk setiap aspek dari data ulasan, sehingga daftar kata-kata lebih relevan. Peneliti mengkodekan kata-kata yang ditemukan lebih dari 100 kali di setiap tipe restoran dari 12 tipe restoran yang ada ke dalam empat aspek tersebut.

Penelitian [8] menunjukkan bahwa makanan adalah aspek yang paling memuaskan pelanggan dengan skor sentimen rata-rata 44%, dan model klasifikasi sentimen yang digunakan cukup baik dalam mengidentifikasi polaritas sentimen untuk setiap aspek. Namun, penelitian [8] juga memiliki beberapa kelemahan, seperti ketergantungan pada daftar kata-kata yang telah ditetapkan sebelumnya. Hal ini dapat membatasi kemampuan model *VADER* dalam menangkap nuansa dan konteks sentimen yang bervariasi. Oleh karena itu, penelitian ini dapat ditingkatkan dengan menambahkan atau mengubah daftar kata-kata yang sesuai dengan domain dan bahasa yang digunakan.

Penelitian lainnya terkait analisis sentimen berbasis aspek adalah penelitian

[17]. Penelitian ini mengkaji sentimen penonton YouTube terhadap ulasan Samsung Galaxy Z Flip 3, yaitu gadget yang populer karena bentuk dan fiturnya yang unik. Penelitian ini menggunakan empat aspek untuk menilai minat (positif) atau ketidaktertarikan (negatif) penonton terhadap gadget tersebut, yaitu 'Desain', 'Harga', 'Spesifikasi', dan 'Citra merek'. Peneliti melabeli dataset secara manual, kemudian mengidentifikasi aspek yang dibahas dalam komentar di kolom YouTube.

Dalam melakukan analisis sentimen berbasis aspek, penelitian [17] menerapkan model *Cross Industry Standard Process for Data Mining* (*CRISP-DM*) dan membandingkan tiga model klasifikasi, yaitu *Naïve Bayes*, *Support Vector Machine* (*SVM*), dan *K-Nearest Neighbor* (*KNN*). Data yang digunakan adalah komentar bahasa Indonesia mengenai Samsung Galaxy Z Flip 3 dari platform YouTube.

Hasil penelitian [17] menunjukkan bahwa mayoritas sentimen komentar adalah positif terhadap aspek desain dan negatif terhadap aspek harga, spesifikasi, dan citra merek. Selain itu, hasil perbandingan model menunjukkan bahwa model klasifikasi *SVM* memiliki kinerja terbaik dengan akurasi keseluruhan sebesar 96.43%, diikuti oleh *Naïve Bayes* dengan akurasi 83.54% dan *KNN* dengan akurasi 59.68%.

Penelitian lain terkait analisis sentimen berbasis aspek adalah penelitian [10], yang melakukan analisis sentimen berdasarkan empat aspek utama film yaitu 'aktor', 'sutradara', 'alur', dan 'musik' menggunakan model *VADER*. Penelitian ini menggunakan data yang sudah memiliki judul, genre, sutradara, aktor, pendapatan, dan rating film yang ditentukan untuk menghindari pengaruh nama film dalam analisis dan mempengaruhi hasilnya karena beberapa nama film mengandung kata-kata sentimen.

Penelitian [10] menunjukkan bahwa model *VADER* mampu menghasilkan akurasi yang baik, kecuali untuk aspek musik. Akurasi keseluruhan sebesar 81%, dengan akurasi masing-masing aspek, yaitu alur sebesar 84%, aktor sebesar 83%, sutradara 80%, dan musik 77%. Namun, penelitian ini memiliki beberapa kelemahan seperti tidak dapat mengenali perbandingan dan informasi tidak relevan dalam komentar, yang tentunya dapat menyulitkan analisis sentimen. Penelitian ini menyarankan untuk menganggap komentar-komentar seperti itu sebagai netral atau

menghapusnya terlebih dahulu agar mendapatkan analisis yang lebih baik.

Penelitian lainnya yang menjadi kajian literatur utama terkait penelitian ini adalah penelitian [18], yang mengembangkan bisnis intelijen dengan menggunakan analisis sentimen berbasis aspek dalam menganalisis ulasan game dan menemukan aspek-aspek yang ada di dalam game. Penelitian [18] menggunakan model *Double Propagation* (*DP*) untuk mengekstrak aspek-aspek yang ada di dalam game dan mengagregasikannya. Aspek-aspek yang diekstrak adalah '*Gameplay*', '*Story*', '*Graphic*', '*Music*', '*Community*', dan '*General/Others*'. Penelitian tersebut menunjukkan bahwa pembuatan bisnis intelijen dengan menggunakan model *DP* pada ekstraksi sentimen aspek menghasilkan *f1-score* 62,39%. Sedangkan agregasi aspek menghasilkan *f1-score* 51,36%. Penelitian ini juga menunjukkan bahwa bisnis intelijen dapat dibangun melalui teks ulasan dengan menggunakan metode ekstraksi sentimen aspek, agregasi aspek, *ETL* (*Extract, Transform, Load)*, dan visualisasi data dengan menghasilkan wawasan dalam bentuk perbandingan antara sentimen positif dan negatif, kategori aspek yang disukai oleh pemain game, dan target konsumen.

Secara ringkas, literatur penelitian-penelitian terkait yang telah dikaji ditunjukkan pada Tabel 2.1.

Tabel 2.1 Studi Literatur

| ***Literatur Review*** | | **Latar Belakang Penelitian** | | **Desain Riset dan Metodologi** | | | |
|---|---|---|---|---|---|---|---|
| **Penulis, Tahun** | **Judul** | **Masalah Penelitian/ Rumusan Masalah** | **Tujuan** | **Metode** | **Data** | **Aspek** | **Hasil/ Temuan/ Kesimpulan** |
| [12] | *Aspect-Based Sentiment Analysis for Hotel Review Using LDA, Semantic Similarity, and BERT* | Analisis sentimen untuk ulasan hotel sering menghadapi masalah ketika data yang diproses tidak berfokus pada aspek-aspek spesifik, sehingga pemilihan istilah dari dokumen ulasan tidak sesuai dengan | Membangun metode otomatis untuk mengategorikan aspek dan sentimen dari ulasan pelanggan pada data skala besar dengan akurasi tinggi. | *Linear Discriminant Analysis* (*LDA), Semantic Similarity, Bidirectional Encoder Representations from Transformers* (*BERT*) | Data ulasan hotel yang berasal dari pengguna dan data yang diambil dari situs web TripAdvisor | 1. *Cleanliness*.<br>2. *Comfort*.<br>3. *Location*.<br>4. *Service*.<br>5. *Food*. | Ekstraksi aspek menggunakan *BERT* dan *Semantic Similarity* menghasilkan performa dengan tingkat presisi sebesar 86% dan *f1-score* sebesar 89%. Sedangkan analisis sentimen menggunakan *BERT* menghasilkan performa tingkat presisi sebesar 96% dan *f1-score* sebesar 97%. Penelitian ini juga menyarankan untuk menggunakan *GPU*, karna model *BERT* menunjukan |

| *Literatur Review* | | Latar Belakang Penelitian | | Desain Riset dan Metodologi | | | |
|---|---|---|---|---|---|---|---|
| **Penulis, Tahun** | **Judul** | **Masalah Penelitian/ Rumusan Masalah** | **Tujuan** | **Metode** | **Data** | **Aspek** | **Hasil/ Temuan/ Kesimpulan** |
| | | tanggapan pelanggan. | | | | | waktu pelatihan yang cukup lama. |
| [13] | *Aspect-Based Sentiment Analysis on Application Review using CNN* | Respons pengguna terhadap kualitas dan layanan aplikasi PeduliLindungi atau SatuSehat masih rendah, dengan rating aplikasi 3.6 dari 5 pada platform *Google Play Store*. | Mengklasifikasikan sentimen pengguna berdasarkan aspek-aspek aplikasi dan memberikan wawasan dan pengetahuan untuk meningkatkan kualitas aplikasi PeduliLindungi. | *Convolutional Neural Network* (*CNN*) | Data ulasan aplikasi PeduliLindungi yang diambil dari *Google Play Store* | 1. *Visual Experience*.<br>2. *Scan – checkin/out*.<br>3. *Vaccine Certificate*.<br>4. *eHac*.<br>5. *COVID-19 Test*.<br>6. *Register / Login*.<br>7. *Performance and Stability*.<br>8. *Privacy, Data, and Security*. | Model *CNN* mampu melakukan klasifikasi sentimen dengan kinerja *f1-score* sebesar 92.23% dalam klasifikasi aspek dan 95.13% dalam klasifikasi sentimen. Sentimen pengguna pada delapan aspek aplikasi didominasi oleh sentimen negatif. |

| *Literatur Review* | | Latar Belakang Penelitian | | Desain Riset dan Metodologi | | | |
|---|---|---|---|---|---|---|---|
| **Penulis, Tahun** | **Judul** | **Masalah Penelitian/ Rumusan Masalah** | **Tujuan** | **Metode** | **Data** | **Aspek** | **Hasil/ Temuan/ Kesimpulan** |
| [14] | Analisis Sentimen berbasis Aspek terhadap Ulasan Hotel Tentrem Yogyakarta menggunakan Algoritma *Random Forest Classifier* | Hotel Tentrem Yogyakarta merupakan salah satu hotel populer bagi para wisatawan yang berkunjung ke Yogyakarta. Namun, dengan banyaknya ulasan hotel yang tersedia di internet, calon pengunjung sering | Melakukan analisis sentimen berbasis aspek terhadap ulasan hotel menggunakan algoritma *Random Forest* untuk mengklasifikasikan polaritas sentimen (positif, negatif, atau netral) terhadap lima aspek utama yang | *Random Forest* | Data ulasan berbahasa Indonesia pada Hotel Tentrem di Yogyakarta melalui situs TripAdvisor | 1. Kamar.<br>2. Pelayanan.<br>3. Lokasi.<br>4. Kolam renang.<br>5. Pusat kebugaran. | Penerapan algoritma *Random Forest Classifier* dalam mengklasifikasikan data ulasan pengunjung hotel dilakukan untuk setiap aspek, namun aspek kamar karena memiliki proporsi sentimen yang seimbang dibanding aspek lainnya. Hasil klasifikasi terhadap aspek kamar yaitu 90% untuk nilai akurasi dan skor f1-nya. |

| *Literatur Review* | | Latar Belakang Penelitian | | Desain Riset dan Metodologi | | | |
|---|---|---|---|---|---|---|---|
| **Penulis, Tahun** | **Judul** | **Masalah Penelitian/ Rumusan Masalah** | **Tujuan** | **Metode** | **Data** | **Aspek** | **Hasil/ Temuan/ Kesimpulan** |
| | | kesulitan untuk menentukan pilihan yang sesuai dengan kebutuhan dan preferensi mereka. | berpengaruh terhadap kualitas hotel, yaitu pelayanan, lokasi, kamar, kolam renang, dan pusat kebugaran. | | | | |
| [15] | *Deep Learning for Aspect-Based Sentiment Analysis on Indonesian Hotels Reviews* | Industri pariwisata tumbuh dengan pesat, didukung oleh kemudahan dalam berbagi pengalaman melalui | Melakukan perbandingan model analisis sentimen berbasis aspek untuk ulasan hotel berbahasa Indonesia. | Delapan model *deep learning* (*RNN, LSTM, GRU, BiLSTM, Attention BiLSTM, CNN, CNN-LSTM, dan CNN-BiLSTM*) | Data ulasan hotel berbahasa Indonesia dari situs web Traveloka | 1. *Price*.<br>2. *Hotel*.<br>3. *Room*.<br>4. *Location*.<br>5. *Service*.<br>6. *Restaurant*. | Dari perbandingan delapan model *deep learning* tersebut, model terbaik untuk klasifikasi aspek adalah model *LSTM* dengan akurasi 92%. Sedangkan model terbaik untuk klasifikasi sentimen adalah *CNN* dengan akurasi 90%. |

| *Literatur Review* | | Latar Belakang Penelitian | | Desain Riset dan Metodologi | | | |
|---|---|---|---|---|---|---|---|
| **Penulis, Tahun** | **Judul** | **Masalah Penelitian/ Rumusan Masalah** | **Tujuan** | **Metode** | **Data** | **Aspek** | **Hasil/ Temuan/ Kesimpulan** |
| | | berbagai platform. | | | | | |
| [16] | Analisis Sentimen Berbasis Aspek pada *Review Female Daily* Menggunakan *TF-IDF* dan *Naive Bayes* | Ulasan-ulasan pada *website Female Daily* perlu dilakukan pengolahan serta analisis agar dapat memberikan informasi atau wawasan dari produk kecantikan untuk produsen maupun konsumen. | Melakukan analisis sentimen berbasis aspek pada ulasan pengguna di web Female Daily yang berbahasa multilingual dengan menggunakan pembobotan *TF-IDF* dan algoritma *Naïve Bayes.* | *TF-IDF, Naïve Bayes* | Data ulasan pelanggan dari web Female Daily | 1. Harga.<br>2. Kemasan.<br>3. Produk.<br>4. Aroma | *TF-IDF* dan algoritma *Naïve Bayes* dilakukan pada tiga skenario. Skenario pertama menghasilkan *F1-Score* sebesar 62,81%. Skenario kedua menghasilkan *F1-Score* sebesar 62,81%, dan skenario ketiga menghasilkan *F1-Score* sebesar 62,81%. |

| *Literatur Review* | | Latar Belakang Penelitian | | Desain Riset dan Metodologi | | | |
|---|---|---|---|---|---|---|---|
| **Penulis, Tahun** | **Judul** | **Masalah Penelitian/ Rumusan Masalah** | **Tujuan** | **Metode** | **Data** | **Aspek** | **Hasil/ Temuan/ Kesimpulan** |
| [17] | Perbandingan *Naïve Bayes*, *SVM*, *dan k-NN* untuk Analisis Sentimen Gadget Berbasis Aspek | Samsung Galaxy Z Flip 3 merupakan salah satu gadget yang sedang marak di kalangan masyarakat karena bentuk dan fiturnya yang unik dan sudah diulas oleh berbagai kanal di platform YouTube | Menganalisis sentimen masyarakat atau penonton di platform YouTube apakah tertarik (positif) atau tidak (negatif) dengan aspek-aspek terkait Samsung Galaxy Z Flip 3 | *Naïve Bayes, Support Vector Machine* (*SVM*), *K-Nearest Neighbors* (*KNN*) | Data komentar bahasa Indonesia mengenai Samsung Galaxy Z Flip 3 dari platform Youtube | 1. Desain.<br>2. Harga.<br>3. Spesifikasi,<br>4. Citra merek. | Dengan model CRISP-DM dan membandingkan model klasifikasi *Naïve Bayes*, *SVM* (*Support Vector Machine*), dan *K-NN* (*K-Nearest Neighbor*), menunjukan mayoritas sentimen adalah positif terhadap aspek desain dan negatif terhadap aspek harga, spesifikasi, dan citra merek. Perbandingan model juga menunjukan model klasifikasi *SVM* memiliki kinerja terbaik dari yang lainnya dengan akurasi keseluruhan sebesar 96.43%. |

| ***Literatur Review*** | | **Latar Belakang Penelitian** | | **Desain Riset dan Metodologi** | | | |
|---|---|---|---|---|---|---|---|
| **Penulis, Tahun** | **Judul** | **Masalah Penelitian/ Rumusan Masalah** | **Tujuan** | **Metode** | **Data** | **Aspek** | **Hasil/ Temuan/ Kesimpulan** |
| [8] | *"How was your meal?" Examining customer experience using Google maps reviews* | Analisis pengalaman pelanggan di industri *hospitality* menjadi sulit ditingkatkan karena adanya ketidaksesuaian antara persepsi pelanggan dengan aspek-aspek pengalaman yang ditawarkan oleh restoran. | Menganalisis pengaruh dari berbagai aspek restoran, seperti harga, kualitas makanan, suasana, dan *value* terhadap kepuasan pelanggan pada berbagai tipe restoran. | *Valence Aware Dictionary for Sentiment Reasoning* (*VADER*) | Data ulasan pelanggan restoran di Google Maps | 1. *Food*.<br>2. *Services*.<br>3. *Atmosphere*.<br>4. *Value*. | Model *VADER* menunjukkan hasil yang cukup baik dalam mengidentifikasi polaritas sentimen untuk setiap aspek yang dibahas dalam kasus ini, dengan makanan sebagai aspek yang paling memuaskan pelanggan dengan skor sentimen rata-rata 44%, diikuti oleh aspek layanan sebesar 25%, aspek suasana sebesar 14%, dan aspek nilai sebesar 10%. |

| ***Literatur Review*** | | **Latar Belakang Penelitian** | | **Desain Riset dan Metodologi** | | | |
|---|---|---|---|---|---|---|---|
| **Penulis, Tahun** | **Judul** | **Masalah Penelitian/ Rumusan Masalah** | **Tujuan** | **Metode** | **Data** | **Aspek** | **Hasil/ Temuan/ Kesimpulan** |
| [10] | *Importance Evaluation of Movie Aspects: Aspect-Based Sentiment Analysis* | Sebagian besar penelitian sebelumnya tentang analisis sentimen terhadap film masih terbatas pada klasifikasi sentimen positif, negatif, dan netral, tanpa menggali secara mendalam aspek-aspek spesifik yang dapat | Mengembangkan analisis sentimen yang lebih holistik terhadap film dengan mengeksplorasi aspek-aspek krusial lebih mendalam, seperti aktor, sutradara, alur cerita, dan elemen musik. | *Valence Aware Dictionary for Sentiment Reasoning* (*VADER*) | Data ulasan film dari web *Internet Movie Database* (*IMDB*) | 1. *Actors*.<br>2. *Directors*.<br>3. *Plot*.<br>4. *Music*. | Model *VADER* mampu melakukan klasifikasi sentimen terhadap aspek-aspek film dengan kinerja secara keseluruhan sebesar 81%, dengan akurasi masing-masing aspek, yaitu aspek alur sebesar 84%, aspek aktor sebesar 83%, aspek sutradara 80%, dan aspek musik 77%. Namun model ini memiliki beberapa kelemahan, yaitu tidak dapat mengenali perbandingan dan informasi tidak relevan dalam komentar, yang dapat menyulitkan analisis sentimen. |

| *Literatur Review* | | Latar Belakang Penelitian | | Desain Riset dan Metodologi | | | |
|---|---|---|---|---|---|---|---|
| **Penulis, Tahun** | **Judul** | **Masalah Penelitian/ Rumusan Masalah** | **Tujuan** | **Metode** | **Data** | **Aspek** | **Hasil/ Temuan/ Kesimpulan** |
| | | memengaruhi penilaian penonton. Oleh karena itu, diperlukan pendekatan analisis sentimen yang lebih rinci, dengan fokus terhadap aspek utama film. | | | | | |
| [18] | *Business Intelligence According to Aspect-Based Sentiment Analysis using* | Di tengah ketatnya persaingan di industri game, pengembang perlu melakukan | Mengembangkan bisnis intelijen dengan menerapkan analisis sentimen | *Double Propagation* (*DP*), *Extract Transform Load* (*ETL*) | Data ulasan pengguna video game pada platform Steam | 1. *Gameplay*.<br>2. *Story*.<br>3. *Graphic*.<br>4. *Music*.<br>5. *Community*.<br>6. *General*/*Others*. | Penerapan bisnis intelijen menggunakan *DP*, *ETL*, dan Tableau berhasil menghasilkan wawasan berupa perbandingan sentimen positif dan negatif, kategori aspek |

| *Literatur Review* | | Latar Belakang Penelitian | | Desain Riset dan Metodologi | | | |
|---|---|---|---|---|---|---|---|
| **Penulis, Tahun** | **Judul** | **Masalah Penelitian/ Rumusan Masalah** | **Tujuan** | **Metode** | **Data** | **Aspek** | **Hasil/ Temuan/ Kesimpulan** |
| | *Double Propagation* | riset pasar secara menyeluruh untuk mengidentifikasi tema, mekanik, pasar, dan tingkat persaingan yang paling relevan sebagai upaya dalam memahami dinamika pasar yang cepat berubah dan menghadirkan | berbasis aspek terhadap ulasan game. | | | | yang disukai oleh pemain game, dan identifikasi target konsumen. Proses ekstraksi sentimen aspek menghasilkan *f1-score* 62%. Sedangkan agregasi aspek menghasilkan *f1-score* 51% |

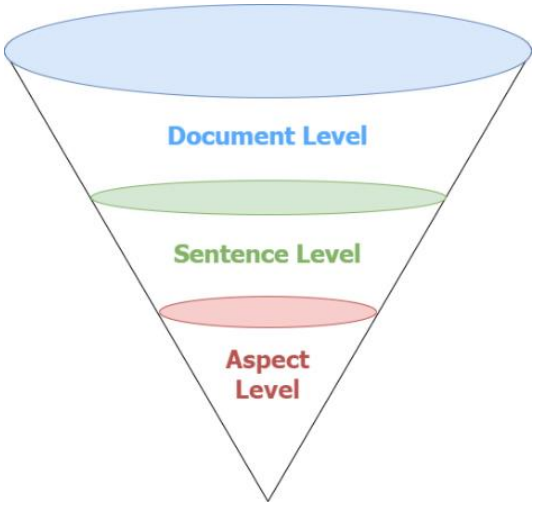
| *Literatur Review* | | Latar Belakang Penelitian | | Desain Riset dan Metodologi | | | |
|---|---|---|---|---|---|---|---|
| **Penulis, Tahun** | **Judul** | **Masalah Penelitian/ Rumusan Masalah** | **Tujuan** | **Metode** | **Data** | **Aspek** | **Hasil/ Temuan/ Kesimpulan** |
| | | produk yang sesuai dengan kebutuhan pasar. | | | | | |

Berdasarkan tinjauan terhadap penelitian-penelitian terdahulu, penelitian ini bertujuan untuk menganalisis sentimen dan aspek-aspek yang mempengaruhi sentimen ulasan pengguna aplikasi Vidio di *Google Play Store*. Penelitian ini menggunakan model *pre-trained IndoBERT* untuk pelabelan sentimen, yang memungkinkan klasifikasi sentimen menjadi lebih efisien. Selain itu, metode *Latent Dirichlet Allocation* (*LDA*) digunakan untuk ekstraksi aspek, yang membantu mengidentifikasi topik-topik utama dalam ulasan pengguna. Data yang digunakan dalam penelitian ini diperoleh melalui proses *scraping*, yang mencakup teks ulasan dan tahun ulasan.

## 2.2 Dasar Teori

### 2.2.1 Opinion Mining

*Opinion Mining* merupakan proses analisis dan ekstraksi opini, sikap, atau emosi yang terkandung dalam teks. Proses ini menggunakan teknik-teknik dari bidang *natural language processing* (*NLP*), *data mining*, dan *information extraction* [19]. *Opinion Mining* bisa diterapkan untuk berbagai macam data teks, misalnya ulasan, buku, artikel, email, atau *website*. *Opinion Mining* bisa digunakan untuk berbagai tujuan, seperti untuk klasifikasi teks, ringkasan dokumen, pemodelan topik, dan analisis sentimen [17,18]. *Opinion Mining* memiliki tiga tingkatan yang berbeda, yaitu tingkat dokumen, tingkat kalimat, dan tingkat aspek [19], yang diilustrasikan seperti Gambar 2.1.




Gambar 2.1 Tingkatan *Opinion Mining*

Tingkat dokumen bertujuan untuk menilai apakah dokumen secara keseluruhan menyampaikan opini yang positif atau negatif. Hal ini didasarkan pada asumsi bahwa dokumen hanya berfokus pada satu objek opini dan memiliki satu sentimen yang dominan [19]. Misalnya, dokumen yang berisi ulasan aplikasi atau film yang memberikan penilaian keseluruhan merupakan contoh dari tingkat dokumen. Selain itu, ada juga tingkat kalimat yang memiliki tujuan berbeda.

*Opinion mining* pada tingkat kalimat bertujuan untuk menentukan apakah sebuah kalimat memiliki opini positif atau negatif. Hal ini didasarkan pada asumsi bahwa setiap kalimat hanya memiliki satu sentimen yang dominan [19]. Misalnya,

kalimat-kalimat yang mengekspresikan pendapat, evaluasi, atau emosi secara langsung merupakan contoh dari tingkat kalimat. Namun, tingkat kalimat tidak cukup untuk menggali opini atau sentimen yang lebih spesifik, sehingga diperlukan tingkat entitas dan aspek.

Tingkat entitas dan aspek bertujuan untuk mengekstrak entitas dan aspek yang menjadi objek opini dan menentukan polaritas opininya. Hal ini didasarkan pada asumsi bahwa suatu kalimat berisi lebih dari satu aspek opini dan sentimen yang berbeda [19]. Misalnya, produk dan fitur-fiturnya, layanan dan kualitasnya, atau orang dan sifat-sifatnya merupakan contoh dari tingkat entitas dan aspek.

### 2.2.2 Analisis Sentimen

Analisis sentimen adalah proses untuk memahami, mengekstrak, dan mengukur emosi atau opini dari teks. Ini sering digunakan untuk mengetahui bagaimana perasaan orang tentang suatu topik, produk, atau layanan berdasarkan data teks seperti ulasan, komentar media sosial, atau artikel berita [22]. Analisis sentimen dapat membantu perusahaan memahami persepsi pelanggan, mengidentifikasi tren pasar, dan membuat keputusan yang lebih baik.

Analisis sentimen biasanya melibatkan beberapa langkah utama:

1) Pengumpulan Data: Mengumpulkan teks dari berbagai sumber seperti media sosial, ulasan produk, survei, dll.
2) Pra-pemrosesan Teks: Membersihkan teks dari elemen yang tidak relevan seperti tanda baca, angka, dan *stopword* (kata-kata umum).
3) Ekstraksi Fitur: Mengubah teks menjadi representasi numerik yang dapat diproses oleh algoritma pembelajaran mesin.
4) Klasifikasi Sentimen: Menentukan polaritas sentimen (positif, negatif, atau netral) dari teks yang telah diproses.

### 2.2.3 *Exploratory Data Analysis* (*EDA*)

*Exploratory Data Analysis* (*EDA*) adalah proses untuk mengamati dan meneliti data, menyimpulkan karakteristik utamanya, dan sering melibatkan metode visualisasi data [21]. *EDA* dapat membantu menetapkan strategi untuk mengolah sumber data sehingga dapat menjawab pertanyaan dan menguji asumsi yang terkait

dengan data. Metode ini juga dapat membantu mengungkap pola, anomali, hubungan, dan struktur data yang mungkin tidak terlihat sebelumnya [23].

*EDA* memanfaatkan berbagai teknik, seperti statistik deskriptif, grafik, plot, dan tabel. Metode ini dapat digunakan untuk memahami dataset, mengeksplorasi hipotesis, dan mempersiapkan data untuk analisis lanjutan [21]. Selain itu, *EDA* juga berperan penting dalam mengidentifikasi dan mempersiapkan data sebelum digunakan dalam model analisis dengan menemukan hubungan, dimensi, kelompok, dan faktor yang melandasi data.

### 2.2.4 *Aspect Based Sentiment Analysis* (*ABSA*)

*Aspect based sentiment analysis* (analisis sentimen berbasis aspek) merupakan pendekatan dari analisis sentimen yang bertujuan untuk mengevaluasi dan mengidentifikasi sentimen pengguna terhadap aspek tertentu dari suatu entitas tertentu. Analisis sentimen berbasis aspek dapat memberikan informasi yang mendalam dan spesifik tentang kekuatan dan kelemahan suatu entitas dari perspektif pengguna.

Analisis sentimen berbasis aspek umumnya melibatkan dua langkah utama, yaitu klasifikasi sentimen dan ekstraksi aspek. Beberapa teknik yang dapat digunakan antara lain adalah pendekatan *rule-based* (berbasis aturan), *deep learning* (pembelajaran mendalam), atau *machine learning* (pembelajaran mesin). Analisis sentimen berbasis aspek dapat diterapkan di berbagai bidang, seperti penilaian produk, film, dan sebagainya.

Analisis sentimen berbasis aspek dapat digunakan untuk menganalisis aspek-aspek layanan, produk, atau fitur dari aplikasi berdasarkan ulasan pengguna [24]. Sebagai ilustrasi, ulasan "*pembayarannya agak ribet cuma bisa pake kredit*" mengindikasikan ketidakpuasan terhadap aspek 'metode pembayaran'.

#### 2.2.3.1 Klasifikasi Sentimen

Klasifikasi sentimen adalah bagian dari analisis sentimen atau *opinion mining* yang fokus pada menentukan polaritas sentimen dari teks. Ini adalah proses untuk mengategorikan teks ke dalam salah satu dari tiga kategori sentimen: positif, negatif, atau netral [19]. Penggunaan tiga kategori sentimen (positif, negatif, dan netral)

dalam analisis sentimen penting untuk memberikan gambaran yang lebih komprehensif tentang sentimen pengguna [11].Sentimen positif dapat menunjukkan aspek-aspek yang disukai pengguna dan dapat dipertahankan atau ditingkatkan. Sentimen negatif mengidentifikasi masalah dan keluhan yang perlu diperbaiki. Sentimen netral memberikan wawasan tentang aspek yang mungkin tidak terlalu mempengaruhi kepuasan pengguna tetapi tetap relevan untuk dipertimbangkan.

Proses klasifikasi sentimen dapat dilakukan dengan berbagai metode, seperti metode *supervised-learning*, metode *lexicon-based* (berbasis leksikon), metode *unsupervised-learning*, atau metode *deep learning*.

1) Metode berbasis leksikon adalah metode yang mengandalkan daftar kata-kata yang sudah diberi label sentimen sebelumnya. Metode ini menghitung skor sentimen dari sebuah teks dengan cara menjumlahkan nilai sentimen dari kata-kata positif dan negatif yang terdapat dalam teks tersebut [25].
2) Metode *unsupervised-learning* adalah metode yang menggunakan algoritma pembelajaran mesin yang tidak membutuhkan data yang sudah diberi label sentimen sebelumnya, tetapi dapat belajar sentimen dari data yang tidak berlabel atau berlabel sebagian. Metode ini biasanya menggunakan teknik klasterisasi atau *semi-supervised* [26].
3) Metode *deep learning* pada klasifikasi sentimen adalah metode yang menggunakan model-model pembelajaran mendalam, seperti *CNN*, *LSTM*, atau *BERT*. Metode ini mempelajari representasi vektor dari teks yang dapat merepresentasikan makna dan konteks dari teks tersebut, dan mengklasifikasikan sentimen berdasarkan representasi vektor tersebut [27].

Setiap metode biasanya menggunakan fitur yang berbeda untuk merepresentasikan teks dan sentimennya. Beberapa fitur yang umum digunakan dalam klasifikasi sentimen adalah:

5) *Terms and Frequency*, yaitu jumlah kemunculan kata-kata tertentu dalam teks yang mencerminkan sentimen penulisnya.
6) *Part of Speech* (*PoS*), yaitu kategori gramatikal dari kata-kata dalam teks yang memengaruhi makna dan sentimen kalimat.
7) *Sentiment Words and Phrases*, yaitu kata-kata atau frasa dalam bahasa yang

mengekspresikan sentimen positif atau negatif.

8) *Rules of Opinions*, yaitu ekspresi atau komposisi bahasa lain yang menyatakan atau mengimplikasikan sentimen dan opini, misalnya kata-kata modal, kata-kata hubung, dan klausa subordinat.
9) *Sentiment Shifters*, yaitu ekspresi yang mengubah orientasi sentimen, misalnya dari positif menjadi negatif atau sebaliknya.
10) *Syntactic Dependency*, yaitu fitur berbasis ketergantungan kata yang dihasilkan dari pohon *parsing* atau pohon ketergantungan.

#### 2.2.3.2 Ekstraksi Aspek

Ekstraksi aspek adalah proses untuk mengidentifikasi entitas target dan aspek-aspek yang berkaitan dengan entitas tersebut dalam suatu dokumen. Entitas target bisa berupa produk, individu, peristiwa, atau organisasi [28]. Proses ini merupakan tahap penting dalam analisis sentimen berbasis aspek, yang bertujuan untuk menemukan objek opini dari ulasan online, misalnya pendapat pengguna tentang berbagai aspek produk yang berbeda, sehingga dapat menganalisis sentimen pengguna secara lebih mendalam dan akurat [29].

Ada berbagai metode yang dapat digunakan untuk mengekstraksi aspek, dan masing-masing metode memiliki pendekatan yang berbeda-beda untuk dapat menghasilkan aspek eksplisit atau aspek implisit. [28]. Aspek eksplisit adalah aspek yang disebutkan secara langsung dalam teks, sedangkan aspek implisit adalah aspek yang tidak disebutkan secara langsung dalam teks, tetapi dapat disimpulkan dari konteks teks [19]. Beberapa metode umum yang sering digunakan adalah:

1) Daftar kata kunci, yaitu metode yang menggunakan daftar kata kunci yang dicocokkan untuk mengekstraksi aspek.
2) *Unsupervised*, yaitu metode yang menggunakan model probabilitas untuk mengekstraksi aspek tanpa memerlukan data pelatihan yang dianotasi. Model ini mengandalkan distribusi probabilitas kata-kata dalam dokumen. Contoh model yang termasuk dalam metode ini adalah *LDA* atau *pLSA*.
3) *Supervised*, yaitu metode yang menggunakan model yang telah dilatih dengan data pelatihan yang dianotasi untuk mengekstraksi aspek. Model ini biasanya

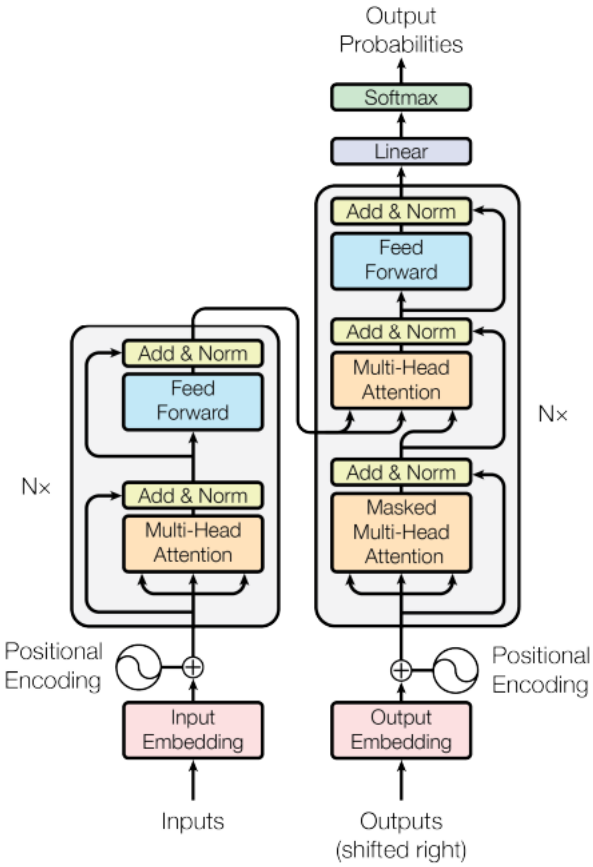
menggunakan algoritma klasifikasi atau pengenal pola. Contoh model yang termasuk dalam metode ini adalah *SVM* atau *CRF*.

4) *Semi-Supervised*, yaitu metode yang merupakan gabungan dari metode *supervised* dan *unsupervised*. Metode ini menggunakan sebagian data yang dianotasi dan sebagian data yang tidak dianotasi untuk mengekstraksi aspek. Contoh model yang termasuk dalam metode ini adalah *Double Propagation* atau *Bootstrapping*.

### 2.2.5 *Bidirectional Encoder Representations from Transformers* (*BERT*)

*Bidirectional Encoder Representations from Transformers* (*BERT*) adalah model *deep learning* representasi Bahasa yang dikembangkan tim peneliti Google untuk menangani berbagai tugas *natural language processing* (*NLP*) seperti *text classification* (klasifikasi teks), *translation* (terjemahan), *summarization* (peringkasan), dan *question answering* (penjawab pertanyaan) [30]. Tujuan dari *BERT* adalah untuk membuat model bahasa yang dapat mempelajari makna dan konteks dari teks secara otomatis, tanpa memerlukan anotasi atau label manual.

*BERT* berbeda dengan model bahasa lainnya, karena menggunakan arsitektur *transformer* yang berbasis teknik *self-attention* dan *masked language modeling*. Teknik *self-attention* membuat *BERT* dapat memperhatikan konteks kata dari kedua arah, sehingga bisa memahami makna dan hubungan antara kata-kata dalam teks [30]. Sedangkan teknik *masked language modeling* membuat *BERT* dapat memprediksi kata yang dihilangkan atau ditutupi dari teks, sehingga dapat belajar struktur dan ketergantungan bahasa [31]. Selain itu, *BERT* dilatih secara *bidirectional* menggunakan data teks yang tidak berlabel dan menggabungkan konteks dari kedua sisi layer [30]. Sehingga, *BERT* bisa disesuaikan dengan menambahkan satu layer saja.

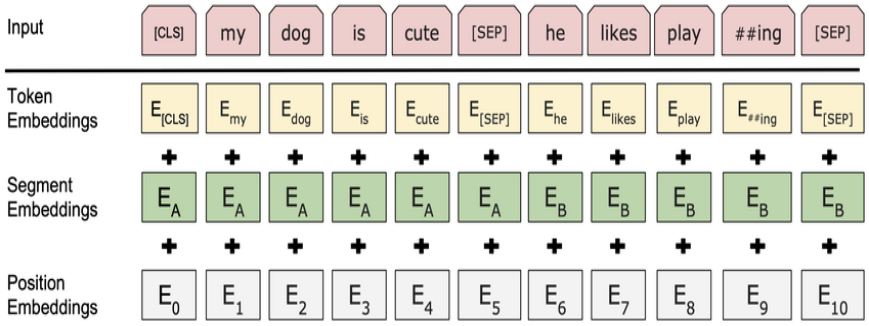



Gambar 2.2 Arsitektur *Transformer* [32]

Arsitektur *BERT* terdiri dari 12 *encoder transformer* yang disusun secara bertumpuk, 12 *attention head*, dan 110 juta parameter [33]. Setiap *encoder transformer* terdiri atas dua sub *layer*, yaitu *feed forward layer dan self-attention layer*. Input untuk *encoder BERT* adalah sebuah urutan token yang diubah menjadi vektor yang kemudian diproses dalam jaringan arsitekturnya [30]. Namun sebelum diproses, *BERT* memerlukan *input* berupa deretan token yang memiliki tiga elemen tambahan, yaitu:

1) *Token embedding*, yaitu elemen dari sebuah token [CLS] yang ditambahkan di awal kalimat dan token [SEP] ditambahkan di akhir setiap kalimat sebagai token input.
2) *Segment embedding*, yaitu elemen penanda yang menunjukkan Kalimat A atau Kalimat B ditambahkan ke setiap token. Ini membantu *encoder* untuk membedakan antara kalimat.
3) *Positional embedding*, yaitu elemen sebuah *embedding* yang posisinya

ditambahkan ke setiap token untuk menunjukkan posisinya dalam kalimat.




Gambar 2.3 Representasi Input *BERT* [30]

Kombinasi dari ketiga *embedding* tersebut merupakan input akhir yang diberikan ke model. Input untuk model BERT dapat berupa sepasang kalimat atau satu kalimat saja. Selain itu, ada dua token unik yang digunakan dalam representasi input, yaitu token *special class classification token* ([CLS]) dan token *separate* ([SEP]) [34]. Token [CLS] adalah token klasifikasi yang digunakan untuk fungsi *next sentence prediction* (*NSP*), token tersebut digunakan untuk membedakan setiap kalimat. Sedangkan token [SEP] adalah token pemisah yang digunakan untuk menandai akhir dari satu kalimat dari satu segmen dan awal dari kalimat lain dari segmen lain [35].

### 2.2.6 *IndoBERT*

*IndoBERT* adalah versi *BERT* yang dilatih kembali dengan data teks Indonesia. Model ini memanfaatkan kerangka kerja dari Hugging Face dan mengadopsi konfigurasi standar untuk *BERT-Base* (*uncased)*. *IndoBERT* hanya dilatih sebagai *masked language model*, yaitu model yang dapat memprediksi kata yang ditutupi dalam sebuah kalimat [36]. Tujuan dari pelatihan ulang ini adalah untuk meningkatkan performa model-model *NLP* Indonesia, karena *IndoBERT* telah menyesuaikan dengan karakteristik bahasa Indonesia, seperti morfologi, sintaksis, dan semantiknya.

Pelatihan model *IndoBERT* menggunakan data *Indo4B*, yang terdiri atas 4 miliar kata bahasa Indonesia dari berbagai jenis dan sumber teks [37]. Dataset ini mencakup teks yang formal, seperti artikel berita dan website, maupun bahasa sehari-hari (*colloquial*), seperti dari media sosial dan blog. *IndoBERT* terdiri dari tiga

sumber utama, yaitu Wikipedia Indonesia, yang berisi 74 juta kata, artikel berita dari Kompas, Tempo, dan Liputan6, yang berjumlah 55 juta kata secara total, dan korpus web Indonesia yang berisi 90 juta kata [36].

Dalam pembuatan *vocabulary (*kumpulan kata), *IndoBERT* menggunakan metode *SentencePiece* dengan *BPE* (*Byte Pair Encoding*) *tokenizer*. *SentencePiece* adalah *subword tokenizer* dan *detokenizer* yang tidak bergantung pada bahasa dan dirancang untuk pemrosesan teks berbasis neural [38], [39]. *Subword tokenizer* adalah metode yang memecahkan kata menjadi unit-unit yang lebih kecil, seperti morfem atau karakter, untuk mengurangi ukuran *vocabulary* dan meningkatkan efisiensi model. *BPE* adalah algoritma yang menggabungkan pasangan *subword* yang paling sering muncul dalam data untuk membentuk *vocabulary* yang optimal.

Model *IndoBERT* menerima input berupa teks ulasan yang disesuaikan dengan representasi input *IndoBERT*. Proses penyesuaian input meliputi langkah-langkah berikut:

1) Mengubah dataset hasil *preprocessing* menjadi token-token kata dengan menggunakan *tokenizer* khusus yang telah disesuaikan dengan bahasa Indonesia dan memisahkan kata-kata berdasarkan spasi dan tanda baca.
2) Menambahkan token khusus seperti token [CLS] di bagian awal kalimat dan token [SEP] di bagian akhir kalimat sebagai pemisah.
3) Menyesuaikan panjang kalimat dengan panjang maksimal input yang diterima oleh *BERT*, yaitu 512 token. dilakukan *padding* dengan menambahkan token [PAD jika kalimat lebih pendek [39]. Dan apabila kalimat lebih panjang dari 512 token, maka dilakukan *truncate* (pemotongan kata).
4) Memisahkan *subword* (sub kata) dengan simbol '##' pada kata yang tidak terdapat di *vocabulary* (*out-of vocabulary*). *Vocabulary* adalah kumpulan kata yang dikenali oleh model *IndoBERT*. Jika ada kata yang tidak ada dalam *vocabulary*, kata tersebut akan dipecah menjadi sub kata yang ada dalam *vocabulary*. Misalnya, jika kata 'paketan' tidak ada dalam *vocabulary*, maka akan dipecah menjadi sub kata 'pa', '##ke', dan '##tan'.
5) Menentukan *attention mask* untuk memisahkan nilai dari token kata dan *padding*. *Attention mask* adalah vektor biner yang menandai mana token yang

merupakan kata dan mana yang merupakan *padding*.

6) Menyiapkan data *loader* untuk setiap data, yaitu data *loader* untuk pelatihan, pengujian, dan validasi. Data *loader* adalah fungsi yang membagi dataset menjadi *batch-batch* yang siap dimasukkan ke dalam model *pre-trained IndoBERT*.

Setelah penyesuaian input dilakukan, token-token tersebut bisa digunakan pada model *pre-trained IndoBERT* menggunakan *layer transformer*, yang terdiri dari *encoder* dan *decoder*. *Encoder IndoBERT* bertugas untuk menghasilkan representasi vektor dari setiap token, yang mencerminkan konteks semantik dan sintaksis dari token tersebut dalam kalimat. Sedangkan *decoder IndoBERT* bertugas untuk menghasilkan output berupa label kelas sentimen (negatif, positif, atau netral)**.**

### 2.2.7 *Latent Dirichlet Allocation* (*LDA*)

*Latent Dirichlet Allocation* (*LDA*) adalah model yang digunakan dalam *topic modelling* (pemodelan topik) dan *topic analysis* (analisis topik) pada data pemodelan berbentuk teks. Model ini mengasumsikan bahwa setiap dokumen dalam *korpus* (kumpulan data teks) dapat direpresentasikan sebagai campuran proporsional dari beberapa *laten* (topik tersembunyi) yang masing-masing memiliki distribusi kata tertentu [40]. Dengan kata lain, *LDA* menganggap bahwa setiap dokumen memiliki beberapa topik yang berbeda, dan setiap kata dalam dokumen berasal dari salah satu topik tersebut.

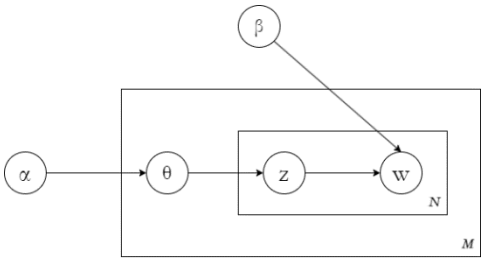
*LDA* dapat digunakan untuk berbagai tujuan, seperti klasterisasi dokumen, peringkasan dokumen, penghubungan dokumen, atau pemrosesan data teks skala besar, karena dapat menghasilkan daftar topik yang berbobot untuk setiap dokumen, dan daftar kata yang berbobot untuk setiap topik [41]. *LDA* juga merupakan model generatif, yaitu model yang dapat memodelkan proses pembentukan dokumen dari topik dan kata. Proses generatif *LDA* untuk setiap dokumen dalam korpus adalah sebagai berikut [40]:

*1)* Memilih topik secara acak dari distribusi *dirichlet* atas topik-topik pada setiap dokumen.

*2)* Memilih kata dari distribusi *multinomial* atas kata-kata yang berkaitan dengan topik terpilih.

*3)* Mengulangi tahapan 1 dan 2 pada seluruh dokumen.

Dengan demikian, *LDA* dapat menghasilkan matriks distribusi topik-dokumen dan matriks distribusi kata-topik, yang dapat digunakan untuk mengeksplorasi dan menginterpretasikan topik-topik yang muncul dalam data teks. Secara ringkas, hubungan antara variabel-variabel dalam proses generatif yang ada pada *LDA* diilustrasikan seperti Gambar 2.2 [40].




Gambar 2.4 Hubungan Variabel dalam Proses Generatif *LDA*

Berdasarkan hubungan antara variabel-variabel yang diilustrasikan pada Gambar 2.2, parameter α dan β diberikan untuk seluruh korpus. Parameter α (*per-document topic distribution*) adalah parameter untuk distribusi *dirichlet* dari topik dalam dokumen, dan β (*per-topic word distribution*) adalah parameter untuk distribusi *multinomial* dari kata dalam topik [40].

Parameter α dan β memiliki pengaruh yang signifikan terhadap hasil analisis topik. Nilai α dan β menentukan seberapa heterogen dokumen dan topik dalam korpus. Nilai α yang besar mengindikasikan bahwa setiap dokumen memiliki proporsi topik yang seragam, sehingga tidak ada topik yang dominan. Sebaliknya, nilai α yang kecil mengindikasikan bahwa setiap dokumen memiliki proporsi topik yang tidak seimbang, sehingga ada satu atau beberapa topik yang menonjol. Sedangkan nilai β yang besar mengindikasikan bahwa setiap topik memiliki proporsi kata yang seragam, sehingga tidak ada kata yang khas untuk topik tersebut. Sebaliknya, nilai β yang kecil mengindikasikan bahwa setiap topik memiliki proporsi kata yang tidak seimbang, sehingga ada satu atau beberapa kata yang khas untuk topik tersebut.

Selain parameter α dan β, terdapat juga variabel θ, Z, dan W yang berada pada tingkat dokumen dan kata. Variabel θ adalah variabel yang merepresentasikan distribusi topik untuk dokumen tertentu. Variabel ini berada pada tingkat dokumen,

yang berjumlah M dalam korpus [40]. Semakin tinggi nilai θ, maka dokumen memiliki lebih banyak topik, sedangkan apabila nilai θ semakin rendah, maka akan semakin spesifik topik dokumen tersebut. Variabel Z dan W adalah variabel ini berada pada tingkat kata yang merepresentasikan topik dan kata dari kata tertentu pada sebuah dokumen [40]. Variabel Z menunjukkan topik yang terpilih untuk kata tersebut, sedangkan variabel W menunjukkan kata yang terpilih dari topik tersebut.

Secara umum, *LDA* bekerja dengan menghitung *joint probability distribution* (distribusi probabilitas bersama), yaitu probabilitas dari suatu kejadian yang terjadi bersamaan, dengan melakukan sampling satu persatu terhadap setiap variabel lainnya. Distribusi probabilitas bersama dapat digunakan untuk mengetahui hubungan antara variabel-variabel yang ada.

Dalam pengelompokan topik, ada dua bentuk distribusi probabilitas yang harus dicari, yaitu distribusi probabilitas dokumen dan distribusi probabilitas topik. Distribusi probabilitas dokumen menunjukkan seberapa besar kemungkinan suatu dokumen mengandung topik tertentu, sedangkan distribusi probabilitas topik menunjukkan seberapa besar kemungkinan suatu topik mengandung kata tertentu [41].

Langkah-langkah penerapan *LDA* dalam pengelompokan topik adalah sebagai berikut:

1) Memuat model dan korpus dari repositori *indolem* atau *indobenchmark* sebagai input untuk *LDA*. Model dan korpus adalah kumpulan data yang berisi dokumen-dokumen yang akan dianalisis.
2) Membuat *dictionary* (kamus) dan korpus dengan mengubah data menjadi bentuk vektor yang terdiri atas kumpulan pasangan kata dan frekuensinya. Frekuensi kata adalah ukuran seberapa sering suatu kata muncul dalam suatu dokumen. Bentuk vektor ini disebut juga *term document frequency* (*TDF*).
3) Mencari *coherence value* (nilai koherensi) sebagai acuan untuk jumlah topik yang optimal dalam pemodelan. Nilai koherensi adalah ukuran seberapa konsisten kata-kata dalam suatu topik berdasarkan frekuensi kemunculan kata-kata tersebut dalam data. Nilai koherensi yang tinggi menunjukkan bahwa topik tersebut relevan dengan data. Skor dari perhitungan koherensi digunakan untuk

mengevaluasi model topik. Semakin tinggi nilai koherensi, semakin baik model yang dihasilkan [42]. Ada beberapa teknik pengukuran koherensi. Dalam studi ini, digunakan teknik pengukuran 'c_v' dengan formula seperti berikut:

$$NPMI(W_i, W_j) = \sum_{j}^{N-1} \frac{\log \frac{p(w_i, w_j)}{p(w_i), p(w_j)}}{-\log p(w_i, w_j)} \qquad (2.1)$$

di mana $p(w_i)$ adalah probabilitas kemunculan acak $w_i$ dalam dokumen, $p(w_i, w_j)$ adalah probabilitas dua kata $w_i$ dan $w_j$ muncul bersama dalam dokumen secara acak. N adalah pilihan tertinggi dari kata-kata $w_1, w_2, \ldots, w_n$ [43].

4) Membangun model *LDA* dengan memasukkan jumlah topik sesuai dengan nilai koherensi yang optimal. Model *LDA* akan menghasilkan distribusi probabilitas dokumen dan distribusi probabilitas topik yang sesuai dengan data.
5) Mencetak kata-kata kunci dengan menampilkan kata-kata yang paling relevan dengan setiap topik beserta bobotnya. Bobot kata menunjukkan seberapa penting kata tersebut dalam topik tersebut. Kata-kata kunci ini dapat digunakan untuk menginterpretasikan makna dari setiap topik.
6) Membuat *similarity matrix* (matriks kesamaan) dengan menghitung nilai kesamaan antara setiap pasangan kata dalam data, berdasarkan *word embeddings* yang dihasilkan oleh model. *Word embeddings* adalah representasi vektor dari kata, yang dapat menggambarkan makna dan hubungan semantik antara kata. Matriks kesamaan adalah matriks yang berisi nilai kesamaan antara setiap pasangan kata, yang dapat digunakan untuk mengukur seberapa mirip dua kata.

Hasil dari *LDA* adalah kelompok-kelompok kata atau klaster yang merepresentasikan topik-topik dalam dokumen. Klaster adalah kumpulan objek yang memiliki kesamaan atau keterkaitan tertentu. Dengan mengetahui distribusi probabilitas dokumen, dapat dilihat bagaimana setiap dokumen terasosiasi dengan berbagai topik dengan tingkat kepercayaan tertentu. Dengan cara ini, juga dapat dilihat bagaimana setiap topik dipengaruhi oleh faktor-faktor lain dalam dokumen.