

Sentiment-Topic Insights

Pengembangan Model Joint Sentiment-Topic untuk Analisis
Sentimen pada Dataset Amazon Review

Disusun oleh:

KELOMPOK 4



Aditya Raja F. K.
2206051626



Amira Shohifa
2206829130



Golda Aurelia S.
2206826173



M. Raffy Zeidan
2206051462



Nabila Parahita
2106726094

BACKGROUND

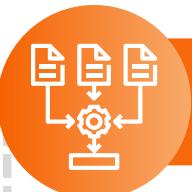
| The Growth of E-Commerce and the Importance of User Reviews



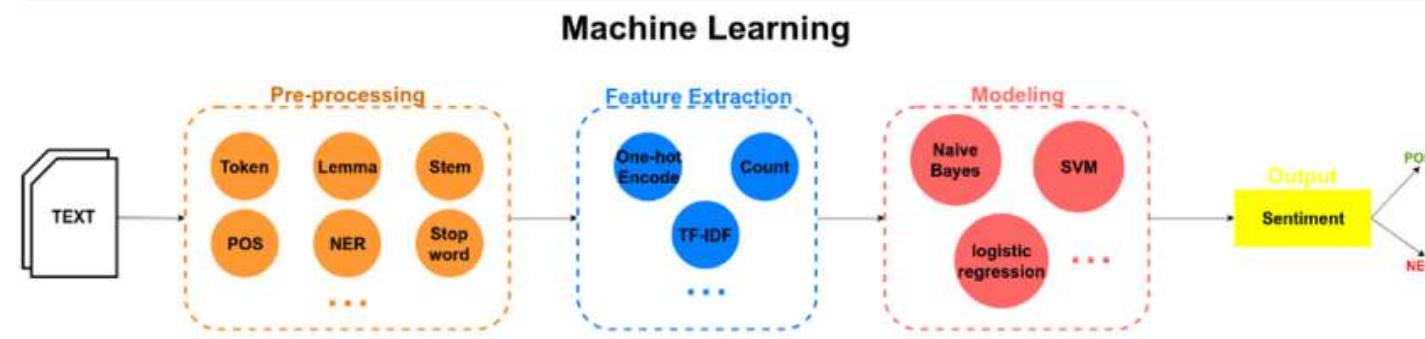
Key Result: Diperlukan model gabungan sentimen dan topik yang efisien, akurat, dan scalable untuk mengolah data ulasan dalam skala besar

RECENT WORKS

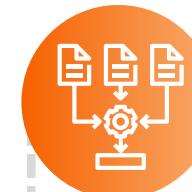
Recent Works of Sentiment-Topic Level Modelling



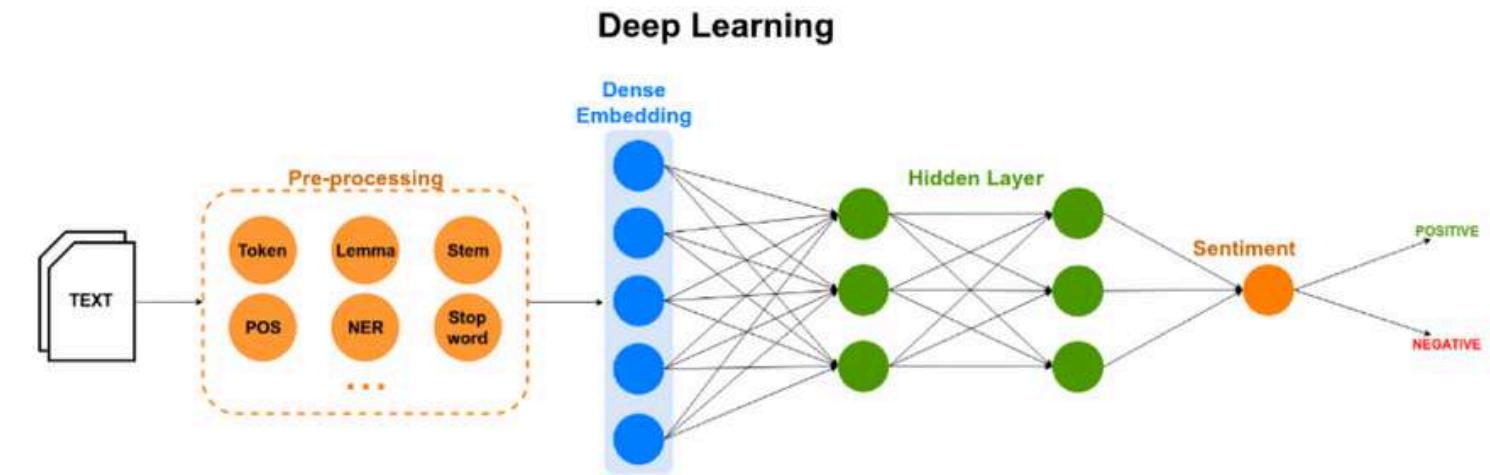
Machine Learning



Merupakan pendekatan awal dalam analisis sentimen yang mengandalkan representasi fitur sederhana seperti Bag-of-Words (BoW) dan TF-IDF. Model seperti Naive Bayes, Support Vector Machine (SVM), dan Logistic Regression digunakan untuk mengklasifikasikan opini berdasarkan fitur tersebut.



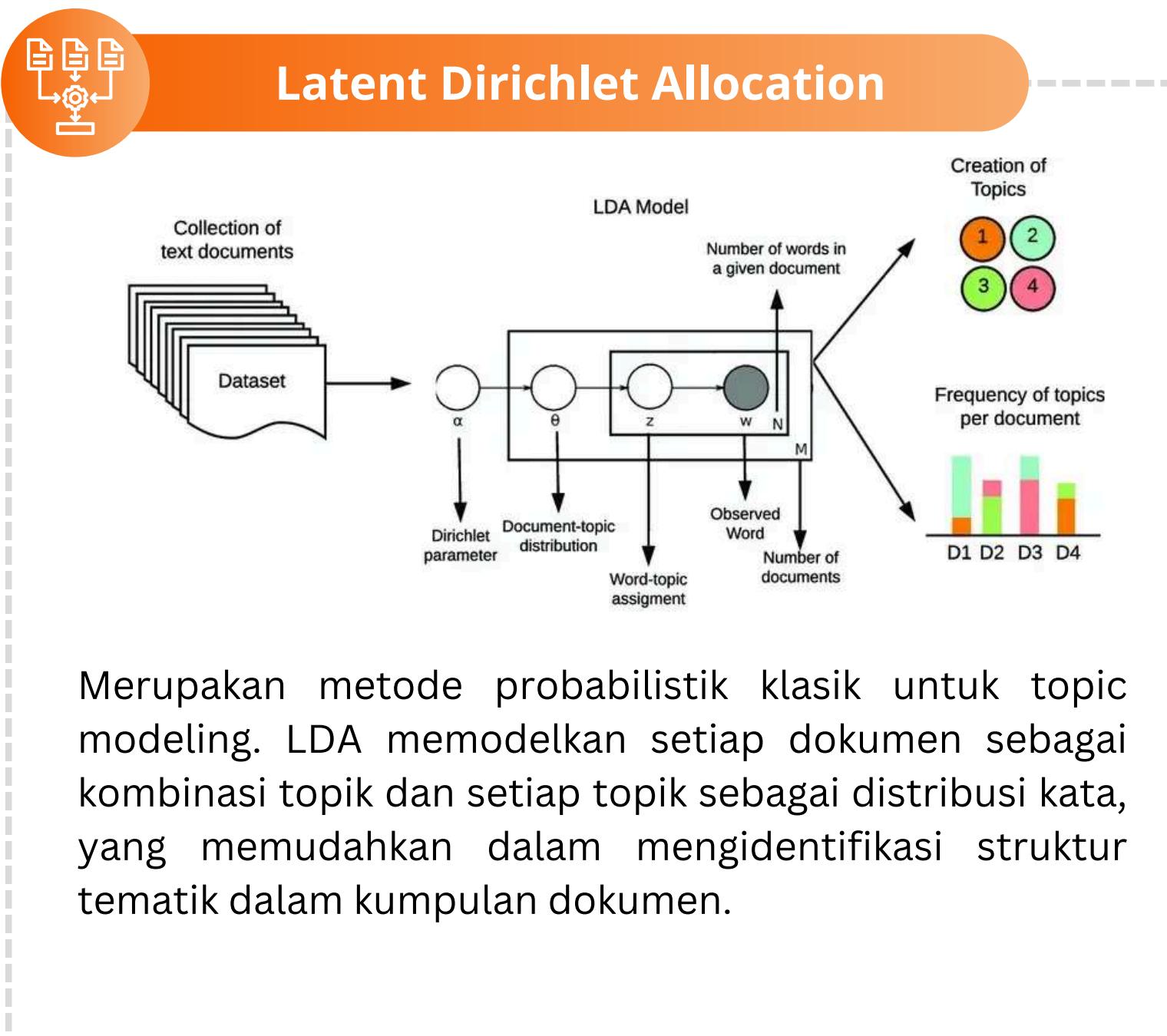
Deep Learning



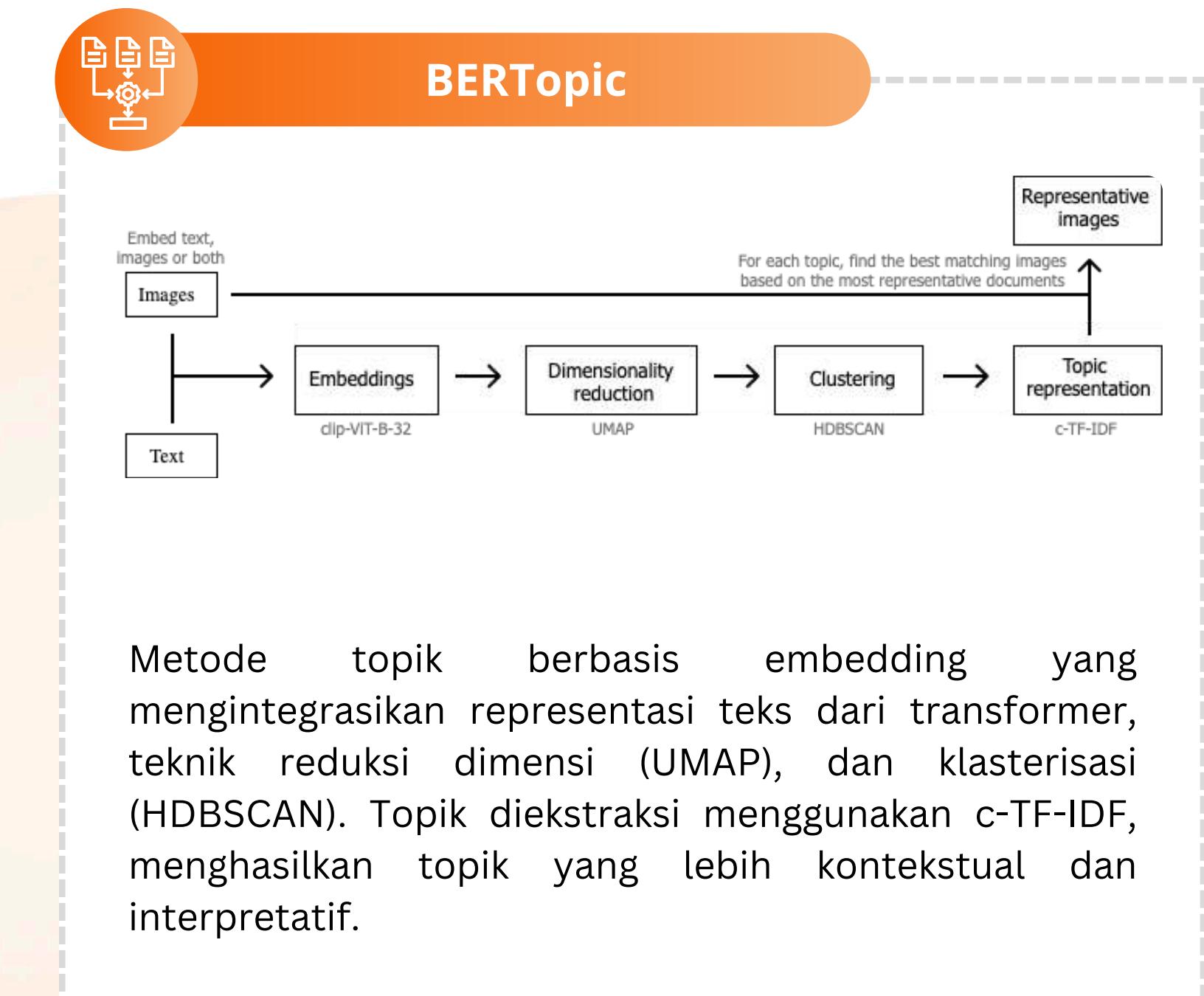
Menggunakan model seperti Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), dan BERT. Pendekatan ini mampu menangkap konteks serta struktur kalimat secara lebih mendalam, sehingga menghasilkan analisis sentimen yang lebih akurat.

RECENT WORKS

Recent Works of Sentiment-Topic Level Modelling



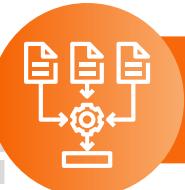
Merupakan metode probabilistik klasik untuk topic modeling. LDA memodelkan setiap dokumen sebagai kombinasi topik dan setiap topik sebagai distribusi kata, yang memudahkan dalam mengidentifikasi struktur tematik dalam kumpulan dokumen.



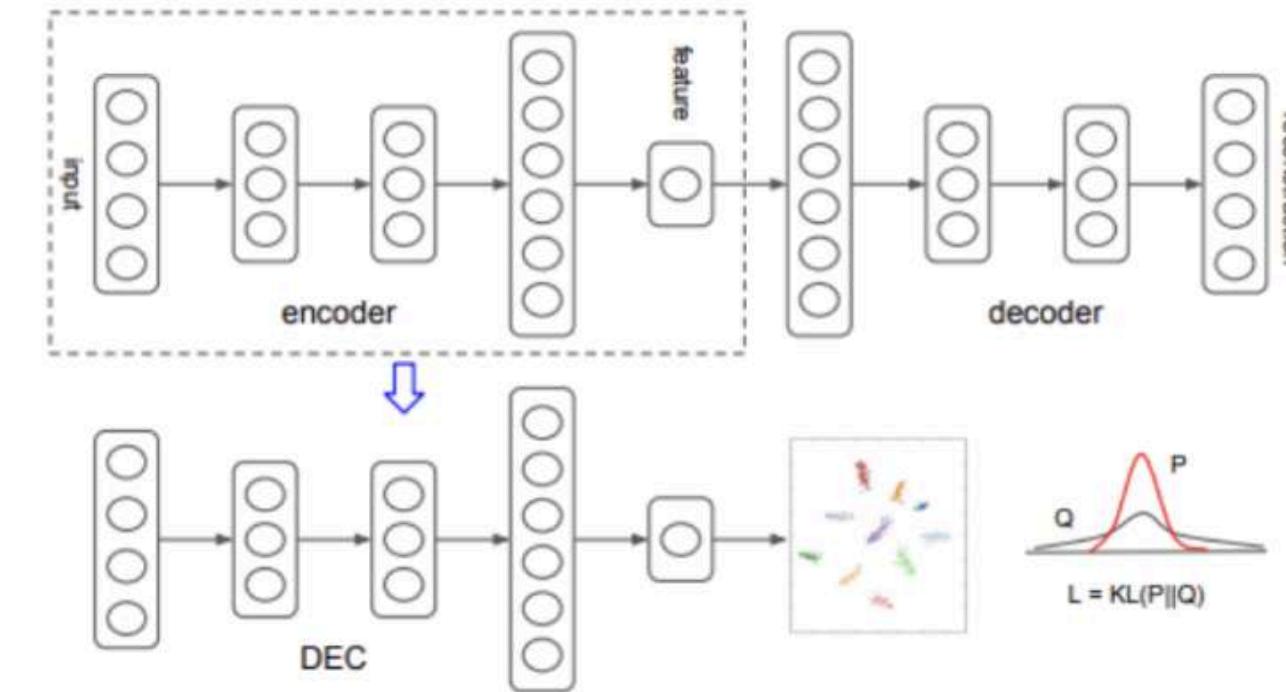
Metode topik berbasis embedding yang mengintegrasikan representasi teks dari transformer, teknik reduksi dimensi (UMAP), dan klasterisasi (HDBSCAN). Topik diekstraksi menggunakan c-TF-IDF, menghasilkan topik yang lebih kontekstual dan interpretatif.

RECENT WORKS

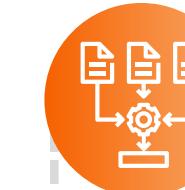
Recent Works of Sentiment-Topic Level Modelling



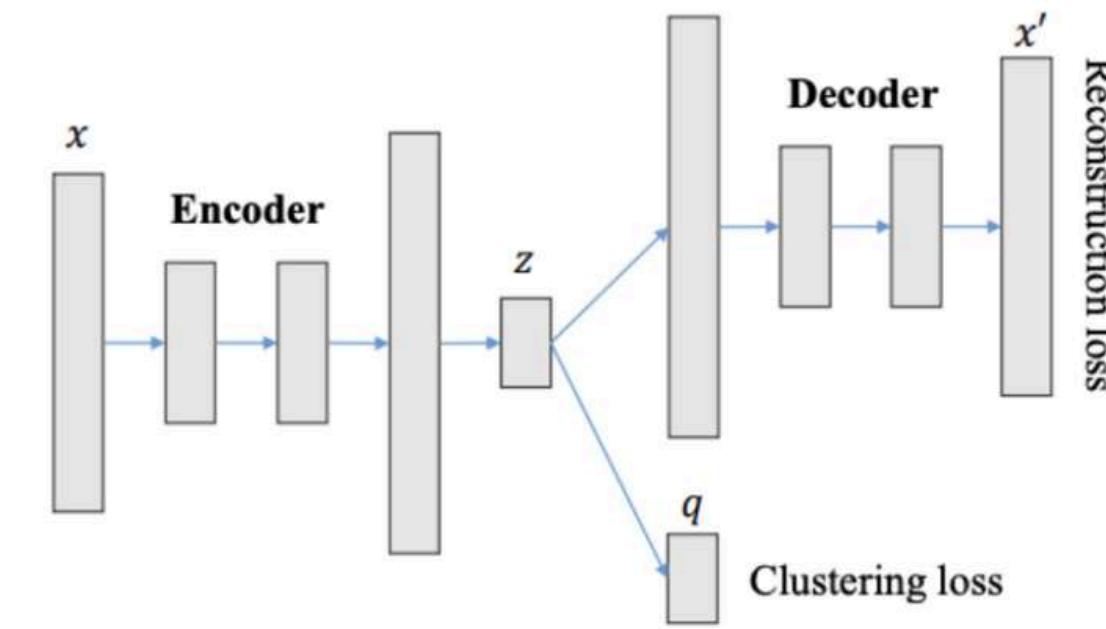
Deep Embedded Clustering



Mengombinasikan autoencoder dengan teknik klasterisasi. Model ini mempelajari representasi data sekaligus membentuk klaster secara bersamaan, menghasilkan pemetaan topik yang lebih terstruktur.



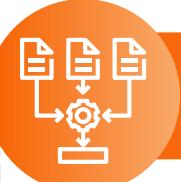
Improved DEC



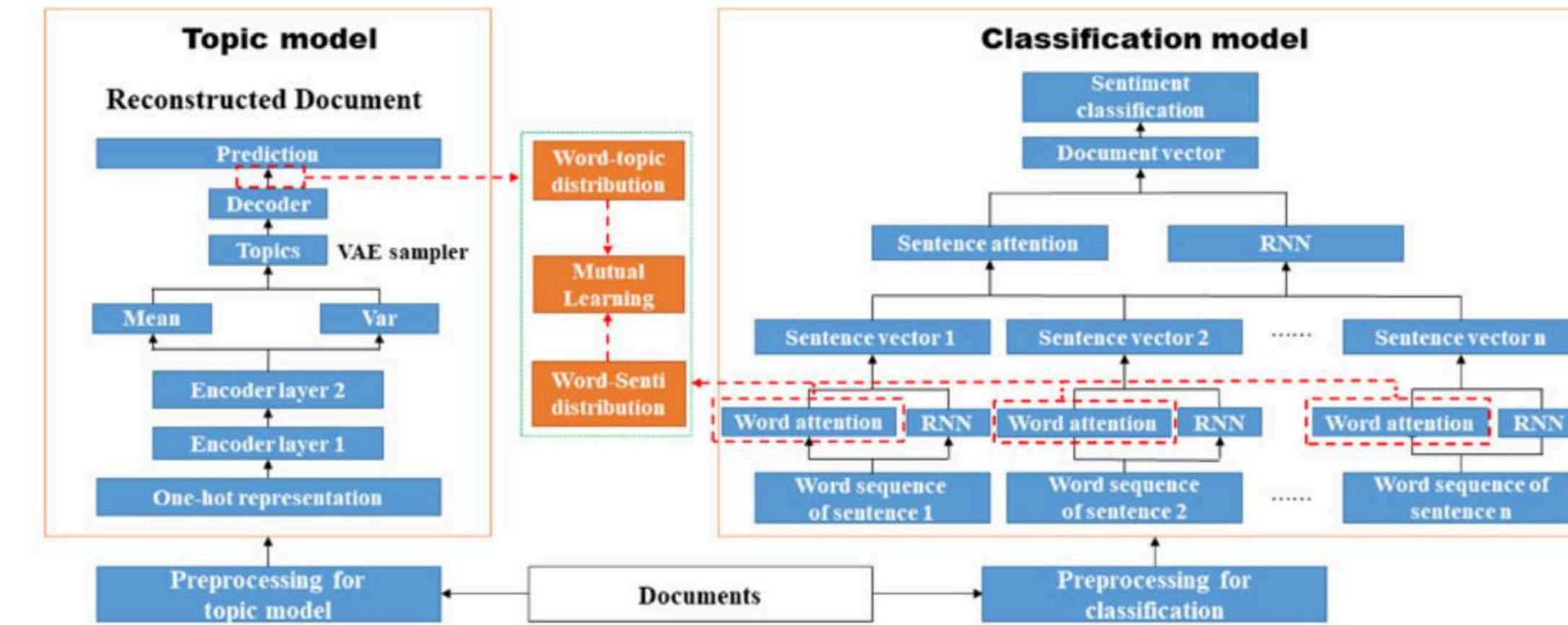
Merupakan pengembangan dari DEC yang mempertahankan struktur lokal data. IDEC menggabungkan reconstruction loss dan clustering loss untuk memperbaiki kualitas hasil klaster.

RECENT WORKS

Recent Works of Sentiment-Topic Level Modelling



Multi-Task Mutual Learning



Merupakan pendekatan multitugas yang secara simultan melakukan klasifikasi sentimen dan deteksi topik. Dengan mekanisme mutual learning, kedua tugas saling memperkuat representasi, menghasilkan peningkatan akurasi dan interpretabilitas.

PROBLEM DEFINITION

General Problem and Problem Statements

General Problem



Klasifikasi sentimen dan pemodelan topik sering dilakukan secara terpisah

Padahal keduanya saling berkaitan



Pendekatan tradisional sering tidak mampu menangkap konteks semantik yang kompleks

Performa model kurang optimal

Key Takeaway: potensi sinergi sentimen dan pemodelan topik terhambat

Problem Statements



Komentar pengguna terlalu banyak untuk dibaca manual

Platform seperti Amazon menghasilkan jutaan komentar setiap hari, sehingga tidak mungkin bagi manusia untuk meninjau semuanya satu per satu.



Perlu model klasifikasi sentimen otomatis

Komentar berisi opini atau emosi pengguna (positif, negatif, netral) yang penting untuk dipetakan secara cepat dan sistematis.



Perlu deteksi topik untuk temukan isu utama

Komentar sering membahas isu spesifik seperti kualitas, harga, atau pengiriman. Topic modeling membantu mengelompokkan dan memahami fokus keluhan/pembahasan.



Gabungan model lebih efisien daripada 2 model terpisah

Joint model menghemat waktu pelatihan, biaya komputasi, dan lebih cocok untuk skenario industri yang membutuhkan kecepatan.



Solusi harus praktis, scalable, dan siap di-deploy

Model yang dikembangkan harus mudah diintegrasikan ke sistem nyata, mampu menangani big data, dan fleksibel untuk berbagai jenis komentar.

PROBLEM DEFINITION

Objectives

General Problem



Klasifikasi sentimen dan pemodelan topik sering dilakukan secara terpisah

Padahal keduanya saling berkaitan



Pendekatan tradisional sering tidak mampu menangkap konteks semantik yang kompleks

Performa model kurang optimal

Key Takeaway: potensi sinergi sentimen dan pemodelan topik terhambat

Objectives

1 **Mengkaji ulang komentar pengguna pada aplikasi Amazon**

2 **Membangun model yang dapat memudahkan analisis sentimen**

3 **Membangun model yang dapat mengidentifikasi topik yang dibicarakan**

4 **Membangun model yang efisien melalui penggabungan analisis sentimen dan topik**

5 **Memberikan solusi yang inovatif, mudah digunakan, dan scalable**

DATA OVERVIEW

Overview of Amazon Product Review Dataset

overall	reviewText
5	It's good for beginners
5	I recommend this starter Ukulele kit. I has everything a beginner needs.
5	G'daughter received this for Christmas present. She loves it.
1	Please pay attention better than I did to the product description.
5	thanx, b
5	Bought as a gift for my daughter and she loved it. The case and tuner are great.

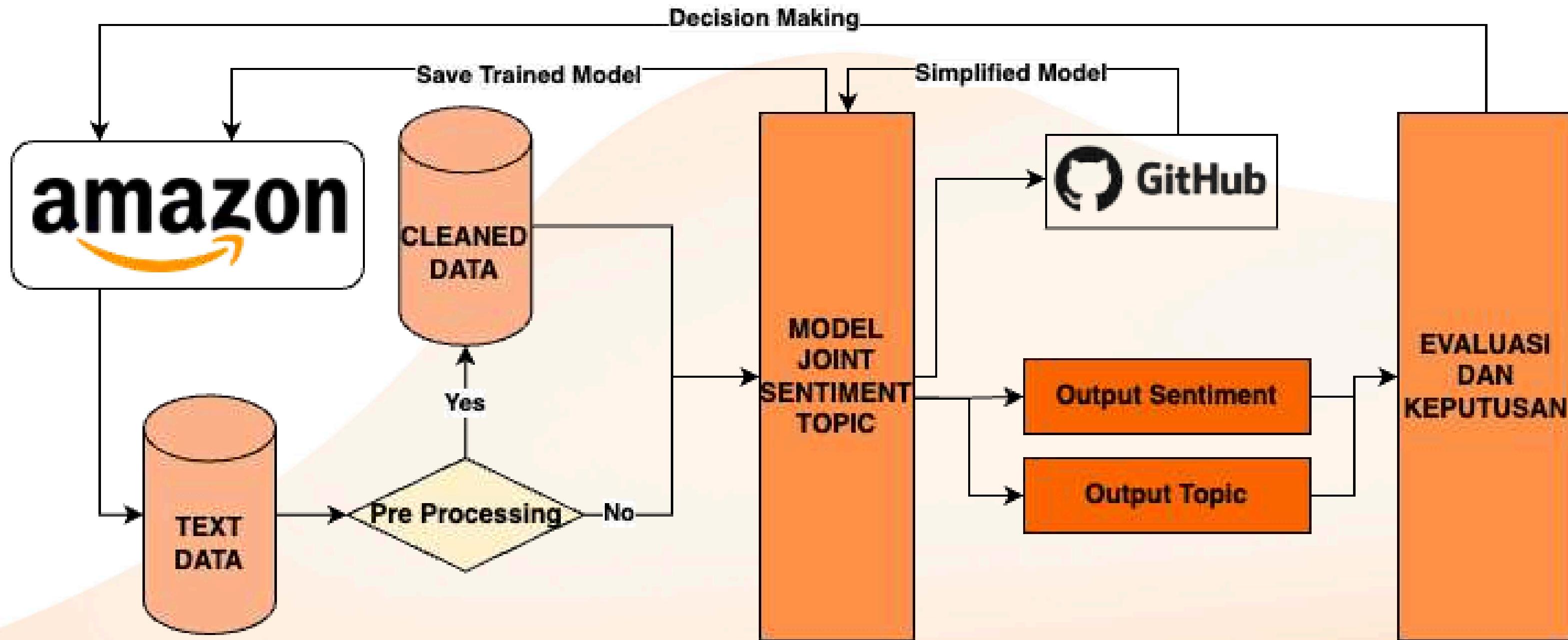
We use a total of 14,000 rows due to GPU limitations

Dataset ini diambil dari kategori **Musical Instruments** dalam dataset **Amazon Product Reviews** yang berisi ulasan dari pengguna Amazon. Dalam penelitian ini, dataset **difilter** untuk hanya mencakup ulasan dengan **rating 1, 2, dan 5 bintang**, dengan fokus pada perbandingan antara ulasan sangat negatif (1-2 bintang) dan sangat positif (5 bintang). Dataset terdiri dari dua kolom utama:

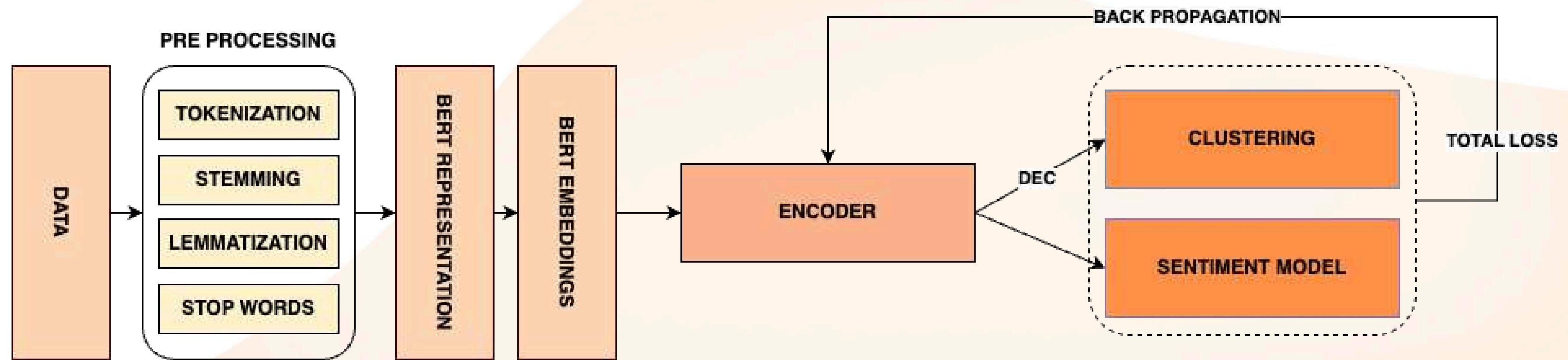
- **overall**: Kolom ini berisi rating bintang yang diberikan oleh pengguna (1, 2, atau 5 bintang).
- **reviewText**: Merupakan teks ulasan yang ditulis oleh pengguna tentang produk yang mereka beli.

Key Result: Menemukan pola sentimen yang kuat dan wawasan pengalaman konsumen di Amazon.

PROPOSED SOLUTION

| **Proposed Solution: Joint Sentiment-Topic Modeling****Solution Workflow**

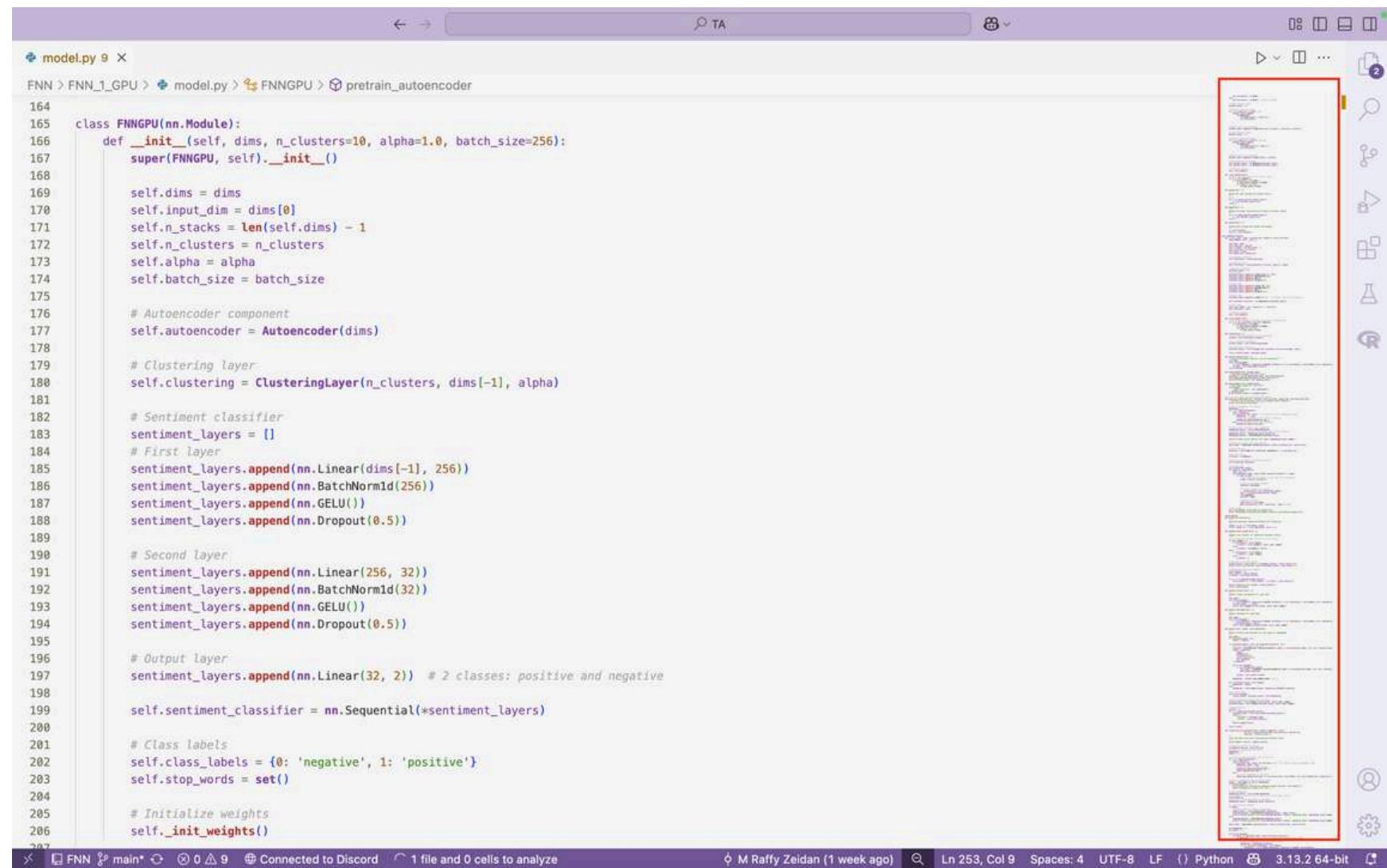
PROPOSED SOLUTION

| **Proposed Solution: Joint Sentiment-Topic Modeling****Proposed Model**

PROPOSED SOLUTION

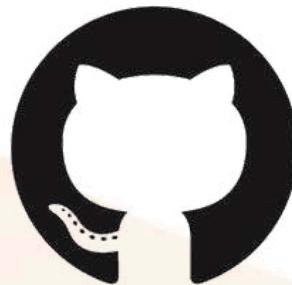
| Simplified Model

Backbone Model



```
model.py
164
165 class FNNGPU(nn.Module):
166     def __init__(self, dims, n_clusters=10, alpha=1.0, batch_size=256):
167         super(FNNGPU, self).__init__()
168
169         self.dims = dims
170         self.input_dim = dims[0]
171         self.n_stacks = len(self.dims) - 1
172         self.n_clusters = n_clusters
173         self.alpha = alpha
174         self.batch_size = batch_size
175
176         # Autoencoder component
177         self.autoencoder = Autoencoder(dims)
178
179         # Clustering layer
180         self.clustering = ClusteringLayer(n_clusters, dims[-1], alpha)
181
182         # Sentiment classifier
183         sentiment_layers = []
184
185         # First layer
186         sentiment_layers.append(nn.Linear(dims[-1], 256))
187         sentiment_layers.append(nn.BatchNorm1d(256))
188         sentiment_layers.append(nn.GELU())
189         sentiment_layers.append(nn.Dropout(0.5))
190
191         # Second layer
192         sentiment_layers.append(nn.Linear(256, 32))
193         sentiment_layers.append(nn.BatchNorm1d(32))
194         sentiment_layers.append(nn.GELU())
195         sentiment_layers.append(nn.Dropout(0.5))
196
197         # Output layer
198         sentiment_layers.append(nn.Linear(32, 2)) # 2 classes: positive and negative
199
200         self.sentiment_classifier = nn.Sequential(*sentiment_layers)
201
202         # Class labels
203         self.class_labels = {0: 'negative', 1: 'positive'}
204         self.stop_words = set()
205
206         # Initialize weights
207         self._init_weights()
```

GitHub Repository



GitHub

Simplified Model

```
sys.path.append('./FNN')
from FNN_1 import FNN, CachedBERTDataset
from FNN_1_GPU import FNNGPU
```

PROPOSED SOLUTION

| Pre-Processing

```

def remove_emojis(text):
    emoji_pattern = re.compile("[ " u"\U0001F600-\U0001F64F" u"\U0001F300-\U0001F5FF" u"\U0001F680-\U0001F6FF" u"\U0001F700-\U0001F77F"
u"\U0001F780-\U0001F7FF" u"\U0001F800-\U0001F8FF" u"\U0001F900-\U0001F9FF" u"\U0001FA00-\U0001FA6F" u"\U0001FA70-\U0001FAFF"
u"\U00002702-\U000027B0" u"\U000024C2-\U0001F251" "]+", flags=re.UNICODE)
    return emoji_pattern.sub(r'', text)

def clean_text(text):
    text = remove_emojis(text)
    text = re.sub(r'^\w\$,.!?', '', text)
    text = re.sub(r'\s+', ' ', text).strip()
    return text

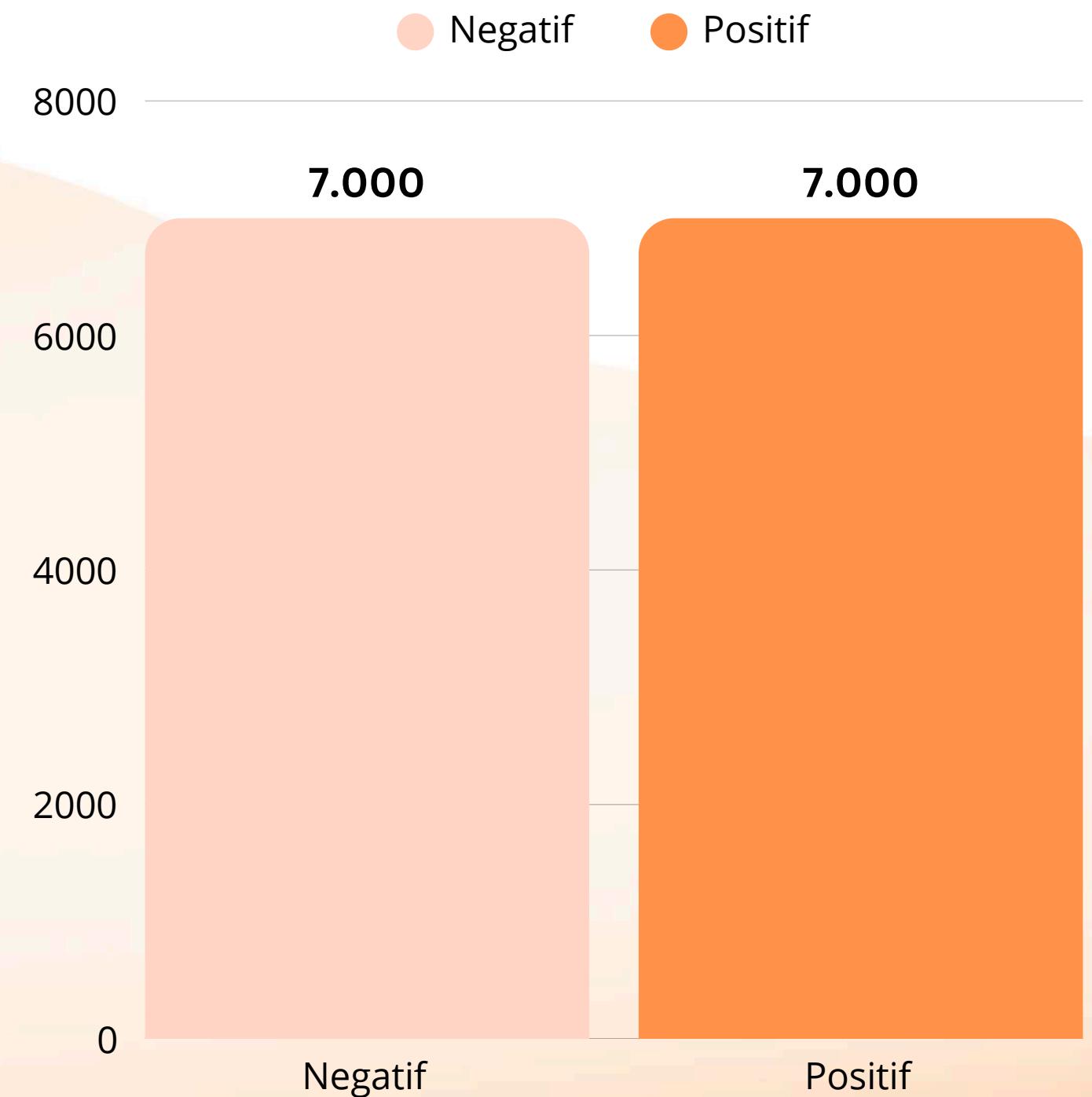
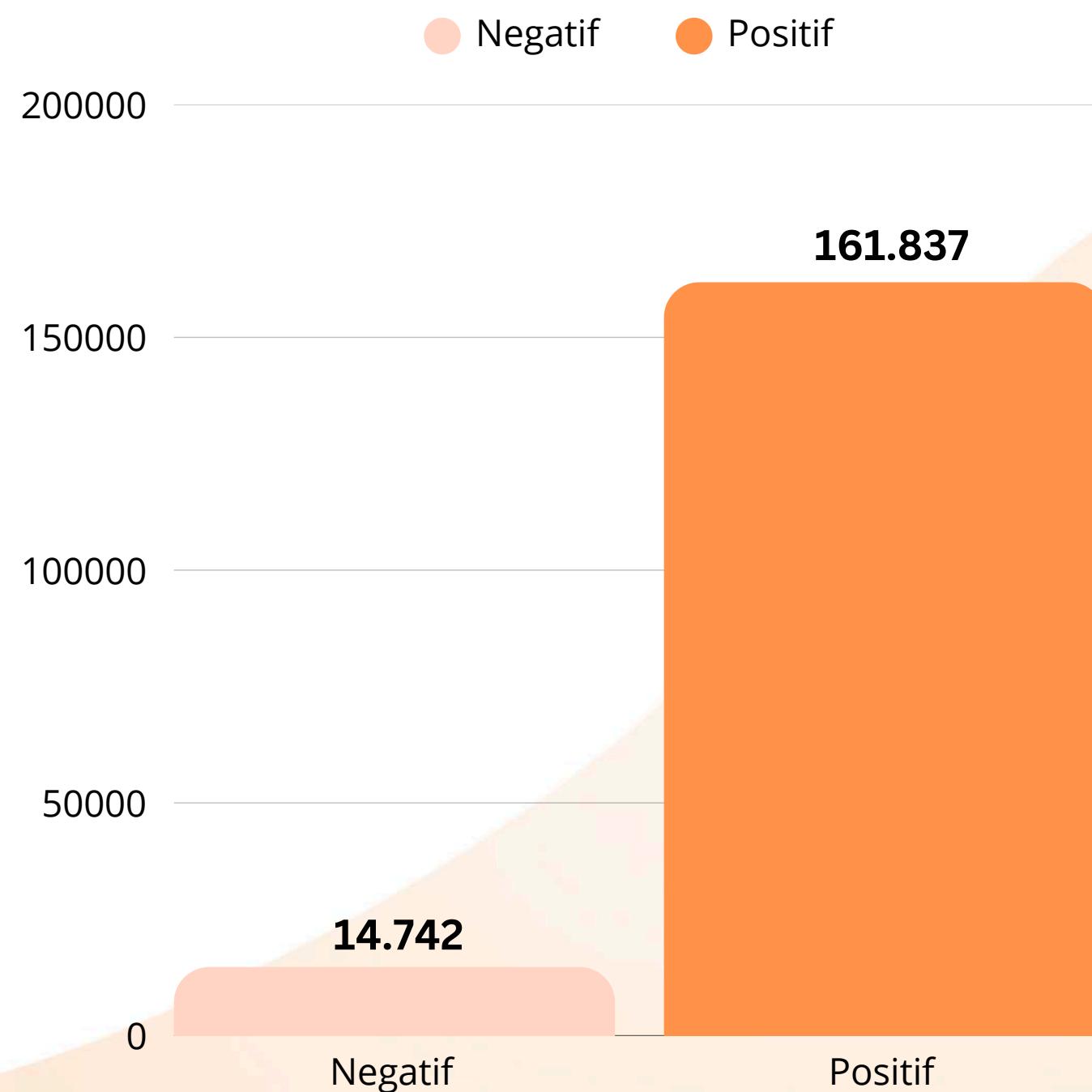
cleaned_data = [clean_text(contents) for contents in tqdm(data['reviewText'], desc="Cleaning Text")]
data['reviewText'] = cleaned_data

```



PROPOSED SOLUTION

| Resampling



PROPOSED SOLUTION

Modelling Schema

1

Model Parameter

- dims = [BERT SIZE , Encoder Layer, Reducted Layer]
- n_clusters = Topic Cluster

2

Autoencoder Parameter

- batch_size = jumlah data per update
- epochs = berapa kali semua data dilatih
- learning_rate = ukuran langkah belajar

3

Training Parameter

- gamma: bobot untuk loss clustering
- eta: bobot untuk loss sentimen
- tol: batas perubahan untuk berhenti (early stopping)
- update_interval: interval menyimpan model
- batch_size: jumlah data tiap step
- maxiter: total langkah training maksimum

```
...  
model = FNNGPU(dims=[768, 500, 500, 2000, 64], n_clusters=4)
```

```
...  
model.pretrain_autoencoder(dataset, batch_size=128, epochs=200)  
model.load_weights('pretrained_ae.weights.pth')
```

```
...  
model.clustering_with_sentiment(  
    dataset, gamma=0.7, eta=1,  
    tol=1e-3, update_interval=140,  
    batch_size=128, maxiter=2e4  
)
```

PROPOSED SOLUTION

| Model Experiment

ML Flow

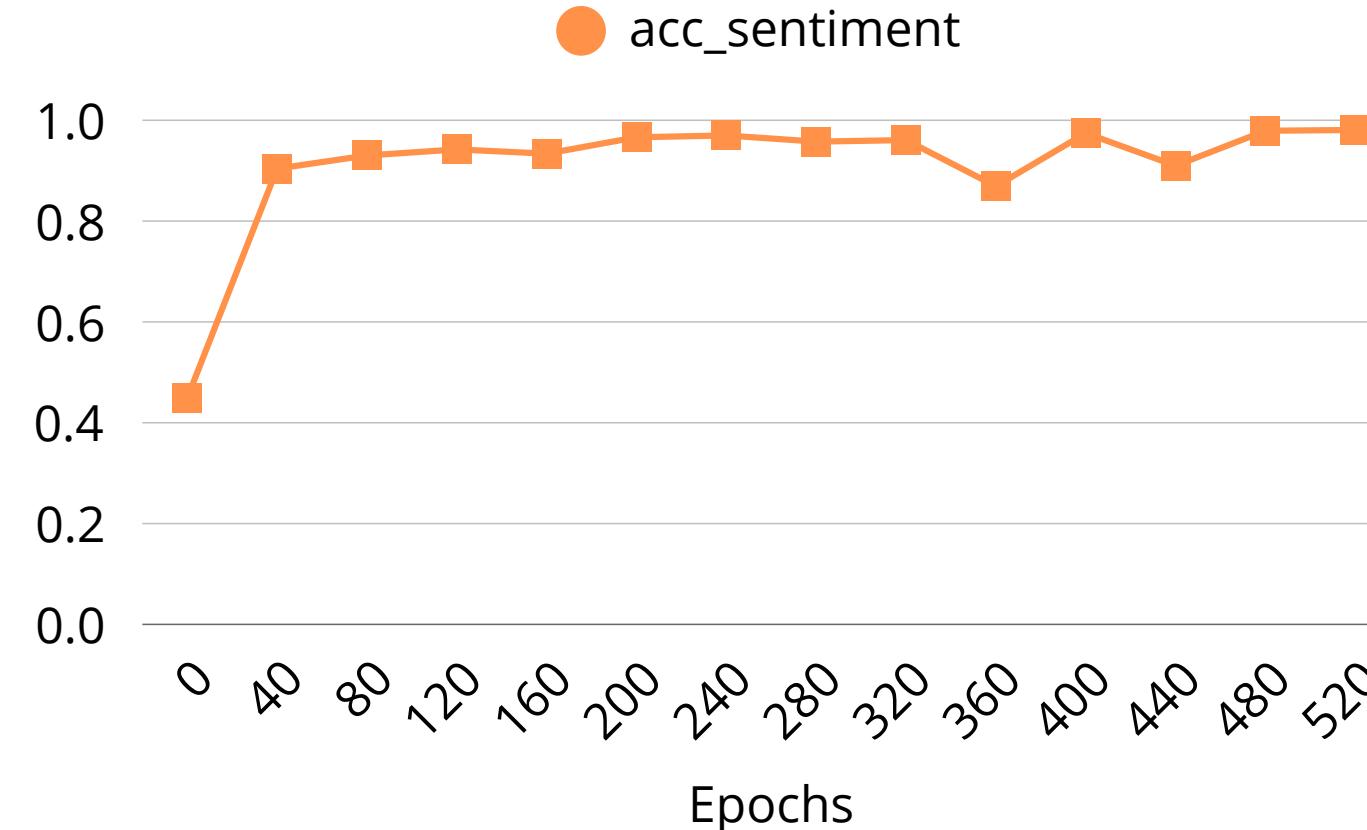
BERT Model	Epoch	Batch	Epoch Encoder	Batch Size Encoder	Total Time	Accuracy
XLM-Roberta	700	128	200	256	640	89.8
Bert-Base-Uncased	700	256	200	128	452	89.29
Bert-Base-Uncased	1000	256	200	128	500	89.6
Bert-Base-Uncased	1000	32	200	128	550	90
XLM-Roberta	700	64	200	128	522	91.43
Bert-Base-Uncased	700	64	200	128	367	90.95
Bert-Base-Uncased	1000	64	200	128	560	86.8
XLM-Roberta	1000	64	200	128	509	90.71

PROPOSED SOLUTION

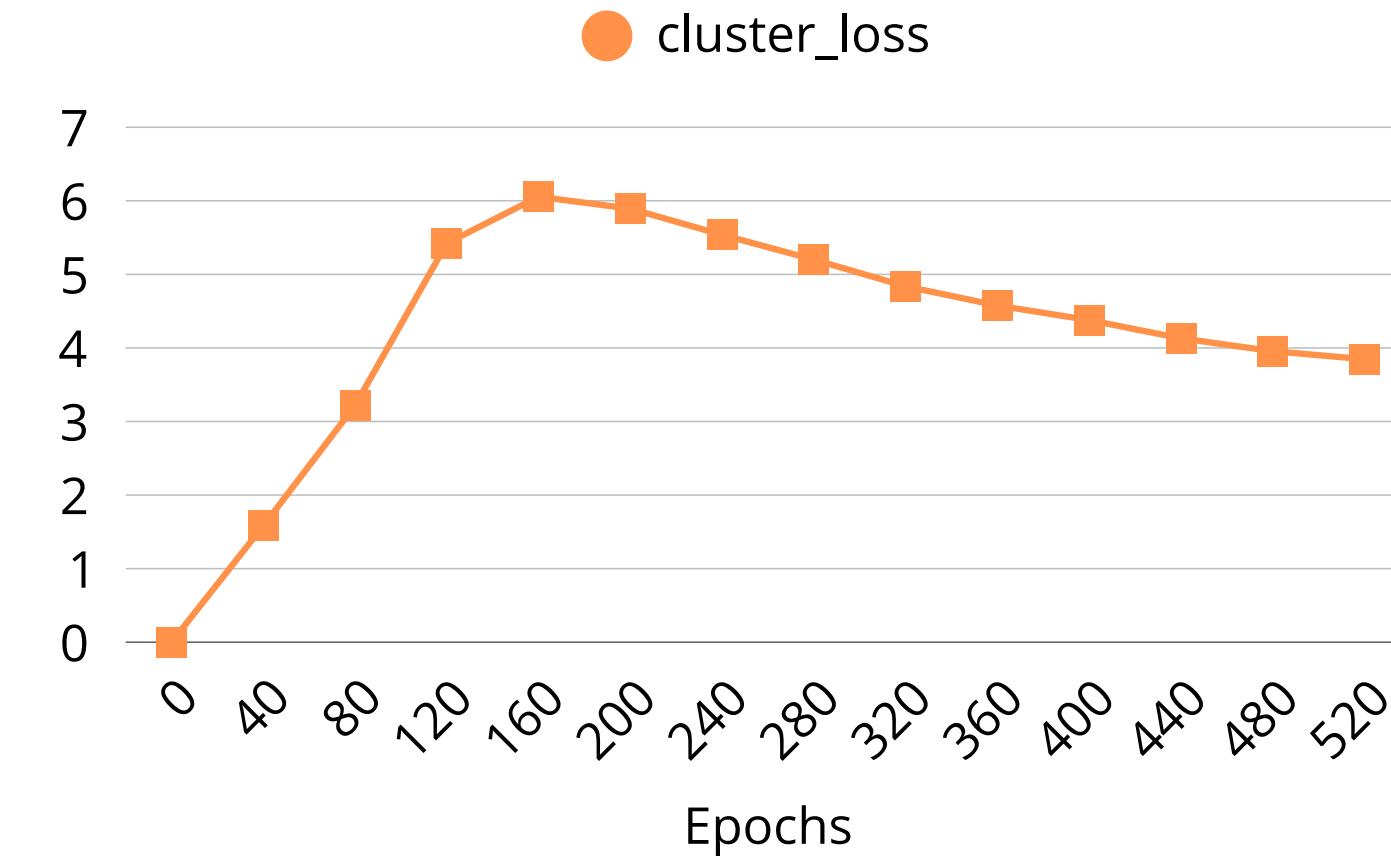
| Model Evaluation

Model Analysis

1 Sentimen Accuracy



2 Clustering Loss



PROPOSED SOLUTION

Model Evaluation

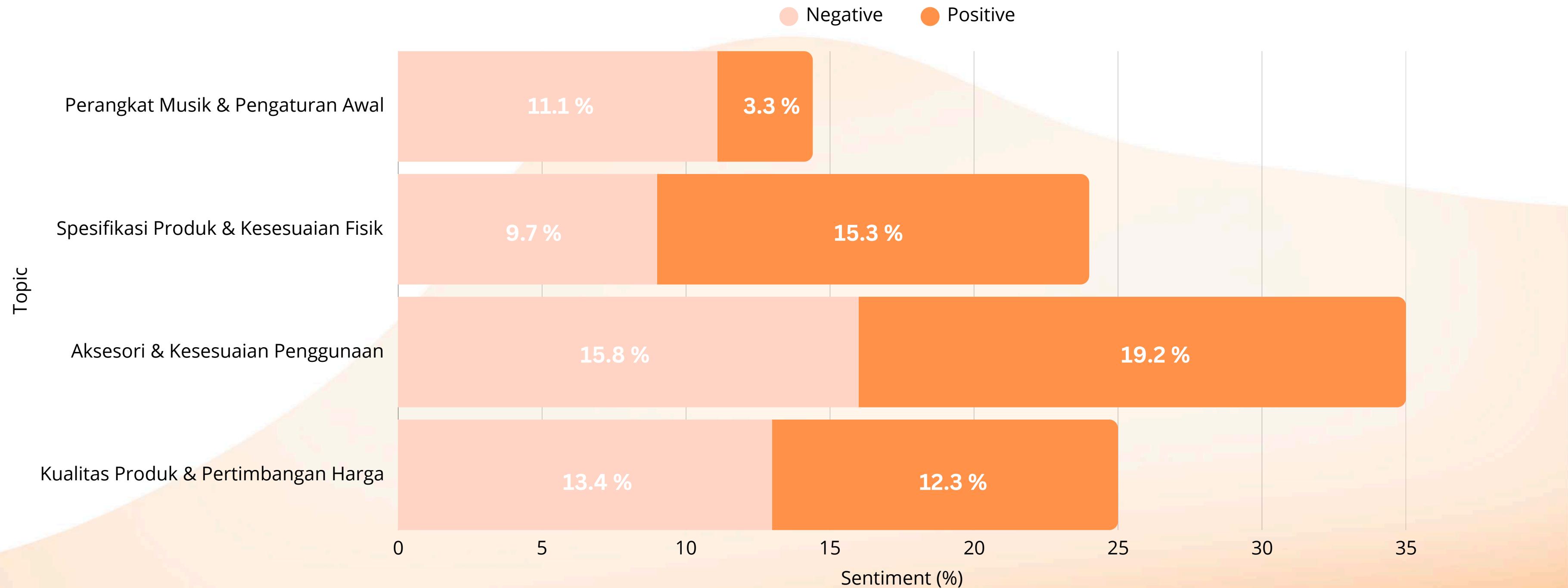
Topic Assign by GPT

Cluster	Common Words
0	quality (377), work (358), time (331), pedal (315), product (281), thing (269), cheap (264), mic (260), price (259), first (258)
1	quality (583), work (475), pedal (429), time (396), easy (384), strap (384), set (383), first (366), recommend (351), price (348)
2	product (181), quality (173), work (134), excellent (127), cheap (91), fit (91), price (77), pedal (74), tone (72), time (69)
3	pedal (304), work (263), quality (259), amp (244), time (230), set (227), first (226), thing (221), put (201), new (195)

cluster :

- 0: "Kualitas Produk & Pertimbangan Harga"
- 1: "Aksesori & Kesesuaian Penggunaan"
- 2: "Spesifikasi Produk & Kesesuaian Fisik"
- 3: "Perangkat Musik & Pengaturan Awal"

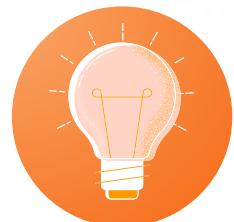
CONCLUSION

| **Summary for Best Models****Result Analysis**

CONCLUSION

| Conclusion Model**Conclusion Model****Pemahaman Kontekstual**

Model yang dikembangkan terbukti dapat menangkap keterkaitan antara konsep sentimen dan topik secara efektif.

**Performa Model**

Model ini berhasil mencapai akurasi yang cukup tinggi, yaitu 91.4%, serta mampu menetapkan topik dengan baik.



15.3 %

Pembuktian Model

Terdapat indikasi bahwa sentimen dan topik dapat direpresentasikan dalam satu konteks semantik yang dapat dipahami, sehingga memungkinkan pembangunan model yang mengintegrasikan keduanya.

CONCLUSION

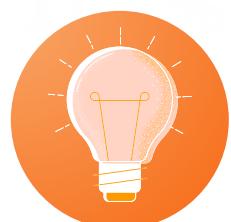
| Conclusion Dataset

Conclusion Amazon



Perangkat Musik & Pengaturan Awal

Didominasi sentimen negatif (11.1%), hanya 3.3% positif → banyak keluhan awal penggunaan.



Spesifikasi Produk & Kesesuaian Fisik

Sentimen positif dominan (15.3%) dibanding negatif (9.7%) → spesifikasi sesuai harapan.



Aksesoris & Kesesuaian Penggunaan

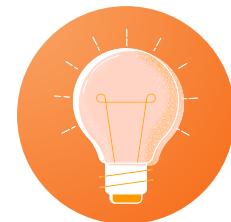
Paling banyak dibahas, dengan sentimen positif tertinggi (19.2%) → aksesoris dinilai sangat mendukung penggunaan.



Kualitas Produk & Pertimbangan Harga

Seimbang, sedikit lebih banyak negatif (13.4%) dibanding positif (12.3%) → pengguna sensitif terhadap kualitas dan harga.

CONCLUSION

| **Recommendation****Overall Recommendation****Perangkat Musik & Pengaturan Awal**

Didominasi sentimen negatif (11.1%), hanya 3.3% positif → banyak keluhan awal penggunaan.

**Spesifikasi Produk & Kesesuaian Fisik**

Sentimen positif dominan (15.3%) dibanding negatif (9.7%) → spesifikasi sesuai harapan.

**Aksesoris & Kesesuaian Penggunaan**

Paling banyak dibahas, dengan sentimen positif tertinggi (19.2%) → aksesoris dinilai sangat mendukung penggunaan.

**Kualitas Produk & Pertimbangan Harga**

Seimbang, sedikit lebih banyak negatif (13.4%) dibanding positif (12.3%) → pengguna sensitif terhadap kualitas dan harga.

Thank You

Topik Khusus II: Web Mining

