

INDONESIAN PRODUCT REVIEW SUMMARIZATION

Richard Dean Tanjaya



PROJECT OUTLINE >>>

01 Dataset Link

02 Project Overview

03 Analysis Process

04 Insight & Findings

05 Conclusion &
Recommendations

06 AI Support Explanation

DATASET LINK >>>>

Dataset diperoleh dari Data in Brief mengenai review product Indonesia yang dipublish pada Oktober 2022

<https://doi.org/10.1016/j.dib.2022.108554>

Attribute	Description
Category	Product classification by category
Product Name	Name of the reviewed product
Location	City name of the shop or product seller
Price	Price in IDR of the reviewed product
Overall Rating	Overall product rating
Number Sold	Total number of products sold
Total Reviews	Total number of reviews given by the customers
Customer Rating	Product rating (range 1 to 5) from the customers
Customer Review	Product reviews given to the product by the customers
Sentiment	Sentiment labels (i.e., Positive, Negative)
Emotion	Emotion labels (i.e., Anger, Fear, Happy, Love, Sadness)

PROJECT >>>> OVERVIEW

Tujuan Project

1

Latar Belakang

2

Permasalahan

3

Pendekatan

4



PROJECT OVERVIEW

Tujuan Project

- **Mengekstrak keywords** dari review produk Indonesia untuk mengidentifikasi tema dan concern utama
- **Menghasilkan ringkasan** yang menangkap esensi dari customer feedback
- **Menyediakan actionable insights** untuk business decision-making melalui sentiment-aware analysis
- **Mendemonstrasikan efektivitas Large Language Models (LLMs)** dalam memproses data teks Indonesia

Latar Belakang & Permasalahan

Dalam e-commerce Indonesia yang berkembang pesat, review pelanggan telah menjadi sumber business intelligence yang kritis. Dengan ribuan review produk yang dihasilkan setiap hari di berbagai kategori, bisnis mengalami kesulitan untuk mengekstrak insight yang actionable dari data teks tidak terstruktur yang sangat besar ini. Analisis manual tradisional memakan waktu dan tidak scalable.

PROJECT OVERVIEW

Pendekatan

Proyek ini menggunakan pendekatan AI-driven modern menggunakan **IBM Granite 3.3-8B Instruct** model untuk natural language processing tasks, yang secara khusus dirancang untuk menangani multilingual text termasuk bahasa Indonesia. Metodologi ini menggabungkan traditional data analysis dengan advanced LLM capabilities untuk mengekstrak deeper insights dari customer reviews.



ANALYSIS PROCESS

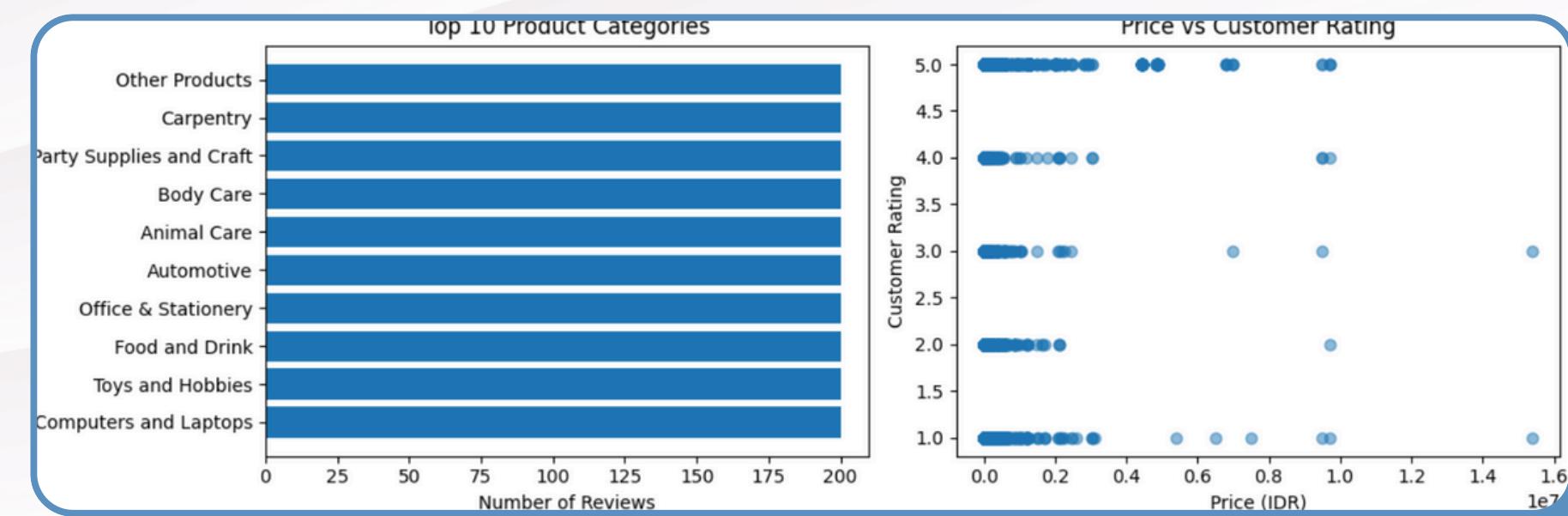
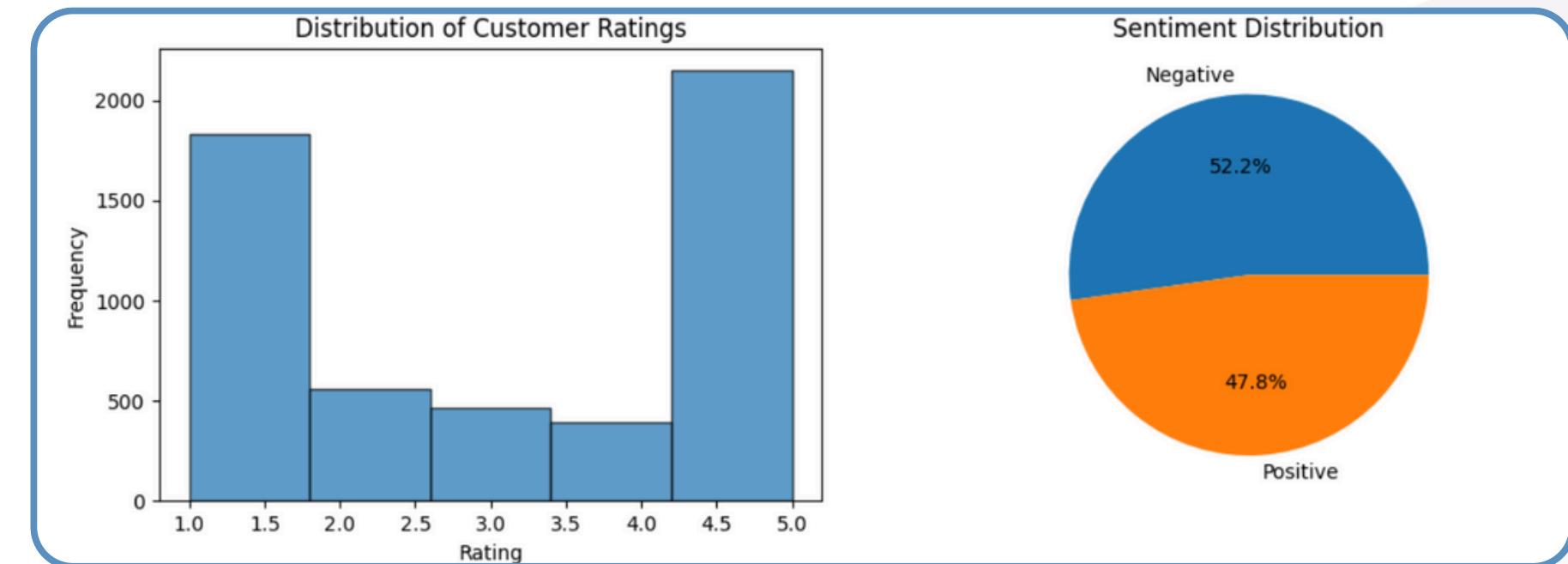
a) Data Preparation dan Exploration

Preparation

- **Size:** 5.400 review dengan 11 kolom
- **Key Features:** Category, Price, Customer Rating, Customer Review, Sentiment

Exploration

- **fig(1):** Customer rating mayoritas terdistribusi pada 1 & 5
- **fig(2):** Sentiment terdapat antara Positive & Negative
- **fig(3):** Product Category tidak imbalance
- **fig(4):** Price VS Rating tidak terlalu terlihat korelasi sama sekali



ANALYSIS PROCESS

b) Data & Text Preprocessing

```
df = (df.groupby('category').sample(n=5, random_state=42).reset_index(drop=True))
display(df.head())
print(f'Rows: {df.shape[0]} x Cols: {df.shape[1]}' )
```

Data compression supaya tidak menghabiskan komputasi yang banyak

```
def text_cleaning(text):
    """
    Minimal cleaning to prepare text for LLM processing:
    1. Handle NaN values
    2. Remove excessive whitespace
    3. Keep original context and nuance intact for better LLM understanding
    """

```

Modern LLMs seperti IBM Granite dapat lebih memahami konteks ketika original structure dipertahankan, oleh karena itu lebih baik tidak perlu Stopwords, Punctuation removing, dll

c) Model Configuration

```
# Set the API token
api_token = userdata.get('api_token')
os.environ["REPLICATE_API_TOKEN"] = api_token

# Model setup
model = "ibm-granite/granite-3.3-8b-instruct"
output = Replicate(
    model=model,
    replicate_api_token=api_token,
)

# Model parameters for consistent results
parameters = {
    "top_k": 50,
    "top_p": 0.9,
    "max_tokens": 200,
    "min_tokens": 10,
    "random_seed": 42,
    "repetition_penalty": 1.1,
}
```

Parameter Optimization:

- top_k=50, top_p=0.9:**
Balanced creativity dan coherence
- max_tokens=200:**
Optimal length untuk summaries
- random_seed=42:**
Reproducible results
- repetition_penalty=1.1:**
Reduced redundancy

ANALYSIS PROCESS

d) Keyword Extraction Process

```
def extract_keywords_with_granite(review_text, sentiment):
    prompt = f"""
    Analyze the following Indonesian product review and extract the MAIN KEYWORDS:

    Review: "{review_text}"
    Sentiment: {sentiment}

    ▲ IMPORTANT:
    - Output **only** a list of keywords – no numbering, no extra text.
    - Output only 1 - 5 keywords, if there's no other keyword that you think is important, then don't include it.
    - Format exactly: keyword1,keyword2,keyword3,keyword4,keyword5
    ....
```

e) Review Summarization Process

```
def summarize_review_with_granite(review_text, rating, sentiment):
    prompt = f"""
    Analyze the following Indonesian product review and write a concise PARAGRAPH summary:

    Review: "{review_text}"
    Rating: {rating}/5
    Sentiment: {sentiment}

    ▲ IMPORTANT:
    - Output **only** one paragraph in Indonesian, consisting of 1-2 sentences.
    - Do not include numbering, bullet points, or any extra text.
    - Keep it factual and focused on the main aspects (product quality, price, service, shipping, etc.).

    Summary:
    ....
```

- **Context-Aware Processing:** Menggabungkan dengan informasi sentiment untuk memandu keyword relevance
- **Output Standardization:** Enforced consistent formatting untuk analysis lebih mudah
- **Quality Control:** Dibatasi hingga 5 keywords paling penting untuk maintain focus

- **Multi-dimensional Input:** Menggunakan review text, rating, dan sentiment label
- **Structured Output:** Consistent paragraph format untuk mempermudah pemahaman output
- **Focus Areas:** Menekankan business-relevant aspects (quality, price, service)

ANALYSIS PROCESS

f) Batch Processing dan Analysis

```
print("\n==== BATCH PROCESSING REVIEWS ===")  
  
sample_df = df.copy()  
  
print(f"Processing reviews for keyword extraction and summarization...")  
  
# Initialize new columns  
sample_df['AI_Keywords'] = ""  
sample_df['AI_Summary'] = ""  
  
# Process each review  
for idx, (index, row) in enumerate(sample_df.iterrows()):  
    if idx % 10 == 0:  
        print(f"Processing review {idx+1}/10...")  
  
    # Extract keywords using original cleaned text  
    keywords = extract_keywords_with_granite(row['Cleaned_Review'], row['Sentiment'])  
    sample_df.at[index, 'AI_Keywords'] = ', '.join(keywords)  
  
    # Generate summary using original cleaned text  
    summary = summarize_review_with_granite(  
        row['Cleaned_Review'],  
        row['Customer Rating'],  
        row['Sentiment'])  
    sample_df.at[index, 'AI_Summary'] = summary  
  
print("Batch processing completed!")
```

- **Sequential Processing:** Menangani setiap review secara individual untuk maintain quality
- **Progress Tracking:** Mengimplementasikan monitoring untuk processing
- **Data Integration:** Mengintegrasikan AI outputs dengan original dataset

KEY INSIGHTS DAN FINDINGS

a) Keyword Analysis Results

```
==== KEYWORD ANALYSIS BY SENTIMENT ====
```

Top 10 Positive Review Keywords:

```
barang: 8  
bagus: 8  
produk: 7  
pengiriman cepat: 6  
sesuai: 6  
aman: 5  
mantap: 4  
terima kasih: 4  
positive: 4  
harga: 3
```

Top 10 Negative Review Keywords:

```
kecewa: 6  
negative sentiment: 5  
beli: 3  
disappointment: 3  
packing: 3  
disappointed: 3  
barang: 3  
rusak: 3  
bahan tipis: 3  
kecil: 2
```

Positive Review:

- **Product Quality Terms:** "bagus", "produk", "packaging", "baik"
- **Service Excellence:** "pengiriman cepat", "mantap", "sesuai", "response", "fast", "sampai"
- **Customer Feedback:** "positive", "terima kasih"

Negative Review:

- **Quality Issues:** "kurang", "rusak", "kecil", "bahan tipis"
- **Customer Feedback:** "kecewa", "disappointment", "issue"

KEY INSIGHTS DAN FINDINGS

b) Summarization Analysis Result

--- Review 4 ---

Original: Barangnya bagus banget, sesuai sama harga, gagangnya juga kokoh banget dan keliatannya kaya bagus gi...

Rating: 5/5

Sentiment: Positive

AI Keywords: barang, bagus, harga, gagang, kokoh

AI Summary: Penjualan produk ini sangat memuaskan, dengan kualitas yang bagus sesuai dengan harga yang diberikan. Gagang produk te

--- Review 5 ---

Original: produk sesuai deskripsi. Recommended ????...

Rating: 5/5

Sentiment: Positive

AI Keywords: produk, sesuai, deskripsi, recommended

AI Summary: Produk tersebut sesuai dengan deskripsi dan diterima dengan rating 5/5, menunjukkan sentimen positif dari pembeli. Pe

Summary Characteristics:

- **Consistency:** Average summary length 1-2 kalimat
- **Completeness:** Berhasil menangkap main review points
- **Language Quality:** Menghasilkan natural Indonesian text dengan grammar yang benar

POSITIVE REVIEWS

NEGATIVE REVIEWS

CONLUSION

Model AI yang dikembangkan dapat dengan mudah mengekstraksi key insights & Summarization terhadap Customer Review. Dapat disimpulkan bahwa **Positive review** paling banyak membahas mengenai kualitas barang dan service seller, sementara **Negative review** lebih membahas mengenai permasalahan kualitas dan feedback customer yang tidak senang.

Keunggulan Sistem:

- **Pengurangan Ketergantungan Tenaga Kerja:** Sistem dapat memproses ribuan review dalam hitungan jam tanpa memerlukan tim analis yang besar
- **Minimalisasi Human Error:** Automated processing menghilangkan inconsistency dan subjektivitas yang sering terjadi dalam manual analysis
- **Standardisasi Output:** Setiap review diproses dengan kriteria yang sama, menghasilkan insight yang consistent dan comparable

RECOMMENDATION

Business Recommendation :

- Seller harus memperbaiki kualitas barang saat dikirim, karena kualitas barang yang burung dapat mempengaruhi kepercayaan pelanggan
- Seller harus bisa meningkatkan kualitas servis, karena mayoritas rating dipengaruhi oleh seberapa baik seller dalam merespon pelanggan
- Realokasi tim analysis dari manual review reading ke strategic insight interpretation untuk lebih fokus pada pemberian solusi

System Recommendation:

- **BERT-model:** untuk contextual understanding yang lebih baik
- **Emotion Detection:** Menambahkan detailed emotion analysis, supaya saat seller merespon customer review dapat disesuaikan dengan kondisi emosi pelanggan

AI SUPPORT EXPLANATION

Proyek ini menggunakan Large Language Model (LLM) untuk 2 primary tasks yang sangat krusial dalam customer review analysis:

1) Intelligent Summarization

- Model memahami full context dari review, termasuk sentiment, rating, dan specific complaints atau praises
- LLM secara intelligent menggabungkan multiple aspects dari review (product quality, service, price, delivery) menjadi summary

2) Advanced Keyword Extraction

- Keywords diekstrak dengan consideration terhadap sentiment context
- Mampu mengidentifikasi compound keywords seperti "pengiriman tepat", bukan hanya single keywords



TERIMA KASIH

