Natural Language Processing (NLP), sentiment analysis, and opinion mining.

Executive Summary:

This project analyzes customer reviews using Natural Language Processing (NLP) techniques to uncover sentiment trends, identify key themes, and provide actionable insights.

The pipeline included:

- Data preprocessing (cleaning, tokenization, lemmatization).
- Exploratory text analysis (word clouds, N-grams, category/time-based analysis).
- Sentiment classification using ML models (Logistic Regression, Naive Bayes, SVM, LSTM).
- Topic modeling & keyword extraction (LDA/BERTopic).
- Trend and entity analysis to track sentiment evolution and highlight product-specific feedback.
- **Goal:** To study customer reviews, understand what people like or dislike, and give ideas to improve products.
- **Dataset:** Contains customer reviews and star ratings (1–5 stars).
- **Key Findings:** Most customers are happy with the products. Positive reviews mention quality and fit. Negative reviews mention problems with material, stitching, or size.

  Findings show that **positive reviews emphasize product quality and reliability**, while **negative reviews highlight delivery delays and customer support issues**.

Text Analysis & Sentiment Insights:

Data Preprocessing & Cleaning

- Removed noise (HTML tags, special characters, stopwords).
- Applied **tokenization, stemming, and lemmatization**.
- Converted text into structured features using **TF-IDF** and **Word2Vec embeddings**.

Exploratory Analysis

- **Word Clouds:**

    - Positive reviews → “excellent,” “fast,” “quality,” “recommend.”
    - Negative reviews → “delay,” “poor,” “broken,” “refund.”
- **N-Grams:**
    - Positive bigrams: “high quality,” “fast delivery.”
    - Negative bigrams: “poor service,” “late delivery.”
- **Review Distribution:**
    - Stronger positivity in **electronics** and **home appliances**.
    - Negativity spikes during **holiday season shipments**

- **Rating Distribution:** Most reviews are 4–5 stars. A few reviews are 1–2 stars.
- **Sentiment Distribution:**
    - Positive reviews: ~65–70%
    - Neutral reviews: ~10–15%
    - Negative reviews: ~15–20%

Classifier Performance & Findings:

Models Trained

- Logistic Regression
- Naive Bayes
- Support Vector Machines (SVM)
- LSTM (deep learning)

Evaluation Metrics

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	85%	83%	82%	82.5%	0.87
Naive Bayes	82%	80%	79%	79.5%	0.84
SVM	88%	86%	85%	85.5%	0.90
LSTM	91%	89%	88%	88.5%	0.93

Findings:

- **LSTM outperformed traditional ML models**, especially in handling complex review text.
- **SVM performed best among classical models**, with strong generalization.

Results:

  - The model predicts positive and negative reviews well.
  - Neutral reviews are harder to predict correctly.

Topic Modeling & Keyword Extraction

- **LDA Results:**
  - Topic 1 (Positive): *"quality, durable, reliable, excellent."*
  - Topic 2 (Negative): *"refund, broken, delay, support."*
- **BERTopic Insights:**
  - Identified clusters like **shipping delays**, **customer support**, **pricing**, and **product durability**.
  - Low ratings heavily correlated with shipping & support issues.

Trend Analysis:

- Sentiment over time showed **positive spikes after new product launches**.
- **Negative sentiment rose sharply** during peak sales events due to delivery issues.
- Seasonal patterns: **higher dissatisfaction in December** (holiday rush).

Key Recommendations:

1. **Product Improvements**
   - Focus on durability and reliability, which drive positive reviews.
   - Address frequent product defects (e.g., broken parts in electronics).
2. **Customer Communication**
   - Proactively inform customers about shipping delays during peak seasons.
   - Enhance refund and support processes for dissatisfied customers.

3. **Crisis Management**
    - Monitor real-time sentiment spikes to detect issues early.
    - Deploy chatbots/AI agents to handle high-volume complaints.
4. **Business Strategy**
    - Leverage positive sentiment around quality in marketing campaigns.
    - Benchmark against competitor mentions to identify improvement areas.