Project Proposal

Large-Scale Sentiment and Topic Analysis of Amazon Product Reviews

Team: AUV
Members: Chenzheng Li, Wenxiang He

Course: CMPT 732 – Big Data Systems   Date: October 20, 2025

Overview

This project focuses on large-scale analysis of the **Amazon Product Reviews** dataset to understand both **micro-level user behaviors** and **macro-level economic trends**. The dataset includes attributes such as `text`, `rating`, `verified_purchase`, `user_id`, and `timestamp`, enabling diverse analyses of user opinions and review dynamics.

At the micro level, we will perform comprehensive sentiment and topic analysis. We will score each review's `text` field using **VADER (from NLTK)** to compute sentiment polarity scores. Based on this, we will:

- Visualize sentiment distributions: *positive*, *neutral*, *negative*;
- Track temporal sentiment trends by month or event (e.g., before/after promotions);
- Compare sentiment across products or brands to detect preference differences.

To extract deeper insights, we will apply **topic modeling** using **LDA** or **BERTopic** to identify dominant themes (e.g., "size", "material", "value for money") and generate automatic product-improvement suggestions. We will also explore **fake review detection** by analyzing `user_id`, `verified_purchase`, and `timestamp` patterns to flag repetitive posting, extreme sentiment values, or possibly generated content.

At the macro level, we will join aggregated annual sentiment results with **World Bank inflation data** (FP.CPI.TOTL.ZG) to examine whether consumer sentiment correlates with global inflation fluctuations. Additional external datasets may be integrated later if interesting correlations are identified.

Methodology

- **ETL (PySpark):** Load and clean JSON reviews, extract metadata (year, country), and compute sentiment and topic scores.
- **Sentiment Analysis:** Compute polarity scores using VADER from the NLTK library.
- **Topic Modeling:** Apply LDA or BERTopic for discovering common review themes.
- **Fake Review Detection:** Identify anomalous users based on high-frequency posting, extreme sentiment, and repetitive review content.
- **Correlation Analysis:** Join yearly sentiment scores with inflation data and compute correlation metrics.

Technologies and Deliverables

We will use **PySpark**, **NLTK**, **Gensim/BERTopic**, **Pandas**, **Matplotlib**, **Seaborn**, and **GitHub**.

Deliverables include:

- A reproducible PySpark pipeline for sentiment/topic analysis;
- Visualizations of trends and distributions;
- A presentation video summarizing key findings and technical insights.

Challenges

Some anticipated challenges include:

- Efficient NLP computation at scale (to be mitigated using distributed Spark jobs);
- Integrating heterogeneous datasets (e.g., Amazon reviews with inflation indicators);
- Ensuring interpretability of black-box models such as BERTopic.