Analyzing Public Discussions for Product Insights

Mining Reddit and Tiki for product issues, sentiment, and trends

Team: Add names and roles here

Objectives

- **Goal:** Mine Reddit & Tiki for product insights.
- **Identify:** Common issues, pros/cons, and sentiment for consumer products.
- **Deliver:** A CLI tool, datasets, and analysis.

Why These Platforms?

- **Reddit:**
  - Rich, threaded discussions.
  - Public API & strong NLP support.
- **Tiki (E-commerce):**
  - Structured, purchase-verified reviews.
  - Complements Reddit with Vietnam market signal.

Data Collection

- **Initial:** PRAW (Python Reddit API Wrapper) for prototyping.
  - **Limitation:** 1,000 post cap, rate limits.
- **Current (Hybrid):**
  - Reddit historical archives (Academic Torrents) to bypass API caps.
  - Tiki review dumps for cross-source validation.
  - **Result:** Broader time windows, more volume.

Subreddits Chosen

- A diverse mix of tech and lifestyle communities:
- r/macbookpro, r/GamingLaptops
- r/iphone, r/AppleWatch, r/Monitors
- r/headphones, r/homelab, r/photography
- …and several others covering home, audio, and PC building.

Preprocessing Pipeline

- **Ingestion:** Ingested JSONL into typed Parquet schemas using Polars & Nushell.
- **Cleaning:**
  - Removed URLs, stripped markup, normalized whitespace.
  - Filtered for English-only content for initial analysis.
- **Modeling Choice:** Modeled each comment individually after experiments showed parent context polluted sentiment signals.

Sentiment Analysis (SA)

- **Approach:** Pretrained transformers over simpler baselines (VADER, TextBlob).
- **Model:** `lxyuan/distilbert-base-multilingual-cased-sentiments-student`
  - **Why:** Good multilingual support, lightweight, strong zero-shot performance.
- **Execution:** Rented GPU on `vast.ai` for large-scale inference.
- **Next Step:** Fine-tune on a domain-specific labeled dataset.

Topic Modeling (Planned)

- **Goal:** Discover key themes and issues per subreddit.
- **Method:** Evaluate BERTopic vs. traditional methods (LDA/NMF).
- **Process:** Preprocess with domain stopwords, lemmatization.
- **Output:** Top topics, representative comments, and trend lines.

CLI Tool (Planned)

- A pipeline for repeatable analysis:
- `ingest`: Raw data to Parquet.
- `clean`: Filter and normalize data.
- `sentiment`: Run batch sentiment analysis.
- `topics`: Train and apply topic models.
- `report`: Aggregate results and export.

Risks & Mitigations

- **Bias:** Sampled randomly across multiple, diverse subreddits.
- **Time Drift:** Included time slices to compare cohorts.
- **Reproducibility:** Pinned environments, config-driven runs, and stored seeds.
- **Ethics:** Used public data, followed platform ToS, aggregated results.

Next Steps

- Finalize EDA visualizations.
- Run sentiment analysis at scale.
- Pilot and select a topic modeling approach.
- Build and demo the core CLI workflow.
- Begin labeling for fine-tuning sentiment model.

Q & A