Signalyze — Ather Rizta Quick Insights Report

Date: 2025-10-15

Executive summary

We collected and analyzed 8,067 English comments about the Ather Rizta from public sources (primarily Reddit and YouTube) to

Top 3 positive highlights (by frequency / supporting examples)
1. App & Software / Usability
   - Many users praised the app features, display integrations and OTA updates; keywords: "app", "update", "display" (positive men
2. Range / Battery Experience
   - Numerous positive comments reference acceptable or better-than-expected range and battery behavior; keywords: "range", "b
3. Ride Comfort & Build (User experience)
   - Users praised ride feel, comfort and perceived build quality across multiple posts; keywords: "ride", "comfort", "build" (positive m

Top 3 areas for improvement (most frequent negative themes)
1. App & Software Stability / Functionality
   - The app and software appear as the single most frequent negative theme as well (~189 negative mentions). Complaints includ
2. Battery / Charging Concerns
   - Negative mentions around battery life and charging (battery ~157, charging ~128) indicate complaints about range or charging
3. Range / Real-world Expectations
   - Users reported range anxiety or lower-than-expected range in some real-world reports (range ~130 negative mentions).

Methodology (short)

- Data sources: `data/reddit_raw.csv` (collected via PRAW) and `data/youtube_raw.csv` (existing). Combined ~8k English comme
- Language detection: `langdetect` to keep only English comments.
- Sentiment: NLTK VADER's SentimentIntensityAnalyzer (fast, zero-cost, interpretable). Label rule: compound >= 0.05 => Positive
- Hierarchical classification: keyword-driven classifier mapping terms to a 3-level taxonomy (Category1 broad area, Category2 cor

Cost-benefit analysis of model choices

- Option A — Transformers (Hugging Face RoBERTa / cardiffnlp):
  - Pros: higher accuracy on nuanced sentiment; supports multilingual and domain adaptation.
  - Cons: requires large model downloads, GPU for reasonable throughput or higher CPU time and cost; slower to iterate.
- Option B — Paid APIs (OpenAI, Google PaLM):
  - Pros: state-of-the-art accuracy, easy to integrate, robust few-shot classification.
  - Cons: ongoing cost per API call for 8k+ records; privacy considerations and rate limits; latency and cost scale with project.
- Option C — Rule-based (VADER) — chosen here for first pass:
  - Pros: Zero direct cost, fast, local, interpretable, sufficient for a high-level triage and theme extraction.
  - Cons: Less accurate on sarcasm, domain-specific terms, and fine-grained aspect sentiment.

Recommendation: For an initial, low-cost insight, VADER + keyword taxonomy is adequate. For a production-ready, high-accuracy

Reproducibility & how to run
---
1. Create a Python 3.9+ environment and install packages in `requirements_out.txt`.
2. Ensure Reddit credentials are set in `.env` (if running PRAW scrape). We wrote to `data/reddit_raw.csv` using the PRAW scrap
3. Run the pipeline:
   - python scripts/analyze_pipeline.py
4. Outputs created:
   - `results/combined_analysis.csv` — combined dataset with sentiment and categories
   - `results/summary.md` — quick stats (also produced)

Limitations

- Dataset bias: majority of data comes from Reddit which may not reflect all buyer demographics.
- Sentiment model: VADER is a general-purpose rule-based model and may mislabel sarcasm or context-specific language; hiera
- Pushshift / YouTube heterogeneity: comment lengths and context differ between platforms; deeper aspect-level models or huma

Next steps (recommended)
---
1. Sample and manually annotate ~500 comments to measure VADER precision/recall and adjust thresholds or train a small class
2. Fine-tune a transformer on the annotated sample for domain accuracy (cost: compute + annotation time). This will improve aspe
3. Build dashboards for weekly monitoring and alerting for spikes in negative themes (service, battery, app outages).

Contact

Signalyze AI & Insights — internal report (deliverables saved in `results/` folder).