PureGym Customer Review Analysis: NLP Topic Modelling Report

CAM_DS_301 Weeks 4-5 | Topic Project 1

Introduction

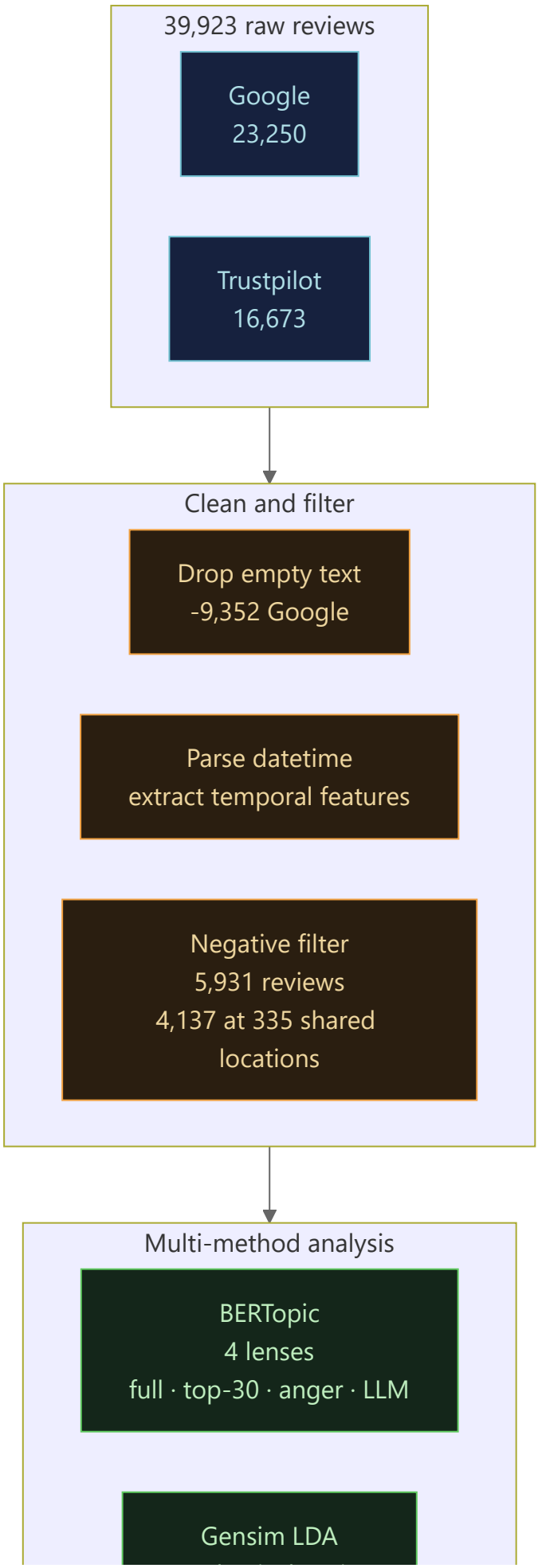
This project analyses 39,923 PureGym customer reviews across two platforms — Google Reviews (23,250) and Trustpilot (16,673) — using a multi-method NLP pipeline: word frequency analysis, BERTopic, Gensim LDA, BERT-based emotion classification, and Qwen2.5-7B-Instruct. The objective is to extract actionable complaint themes that PureGym management could use to reduce negative reviews and improve customer retention.

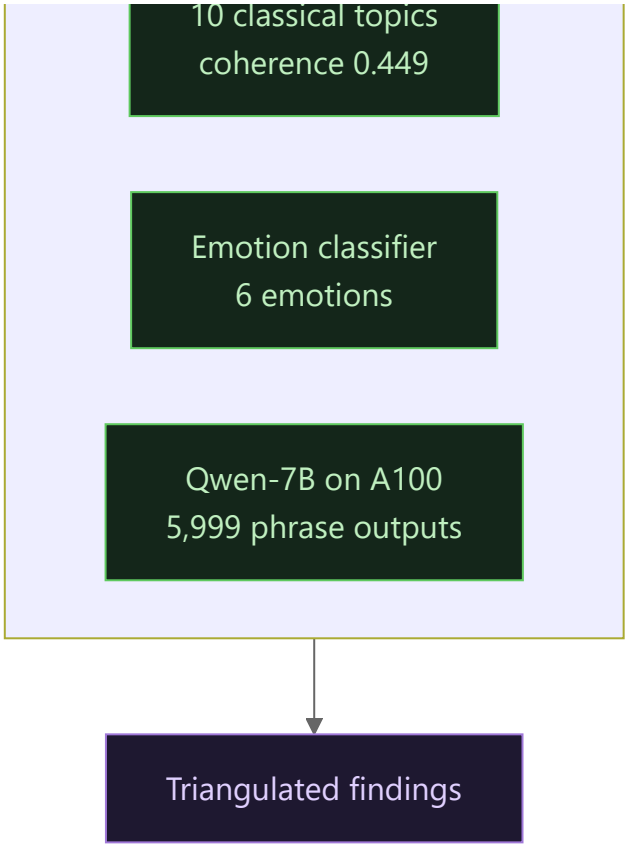
Methodology

The analysis followed a progressive experimentation approach. After loading and cleaning the data (removing 9,352 Google reviews with no text), we parsed datetime fields and filtered to 5,931 negative reviews (score < 3); 4,137 of those came from 335 locations common to both platforms (after manual cross-platform name-merging) and fed cross-platform topic modelling.

Text preprocessing was evaluated using a workbench of 10 configurations. A critical finding: **heavy preprocessing hurts BERTopic**. Lemmatization increased outliers from 36.7% to 47.6%, because BERT embeddings rely on natural language context. Zipf's Law analysis (slope -1.034, R^2=0.993) confirmed why: BERT was trained on the full Zipf distribution, so stripping stopwords strips signal. The optimal pipeline feeds raw text for embeddings while using a CountVectorizer with custom stopwords and bigrams for labels.

BERTopic was applied as four complementary lenses: full negative reviews at common locations, the top-30-locations subset, anger-filtered reviews, and an LLM-driven run where Qwen2.5-7B-Instruct first extracted natural-language topic phrases from each anger-filtered review and BERTopic then meta-clustered the resulting 5,999 phrase outputs. UMAP is seeded (`random_state=42`) across all four runs for reproducibility. Gensim LDA ran on the same negative pool as a classical comparison (10 topics, coherence 0.449). Emotion analysis classified every negative review across six categories. Qwen2.5-7B-Instruct substitutes for the rubric's Falcon-7B (rationale in Appendix); a per-method comparison table is in the Appendix.

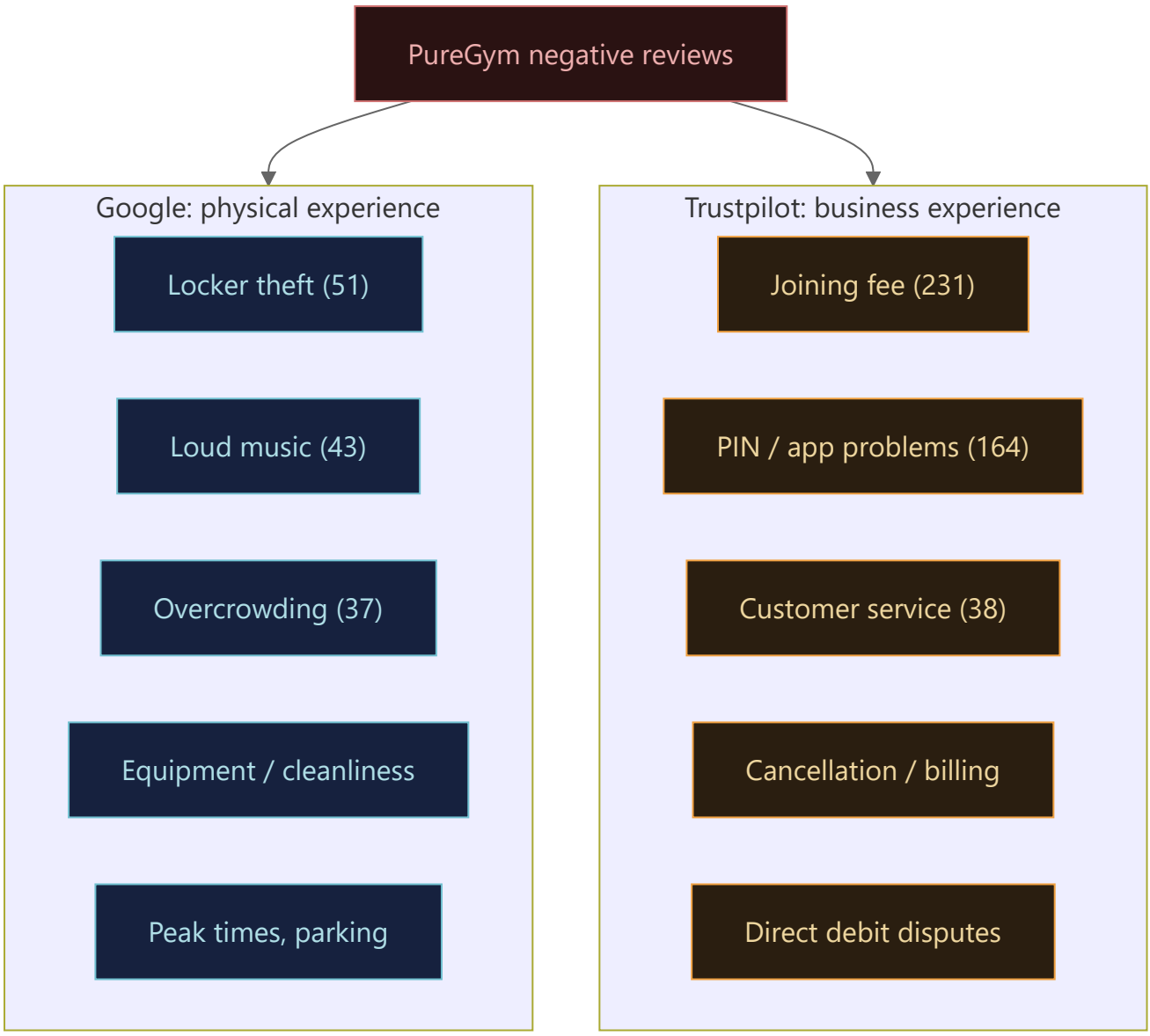

39,923 raw reviews
Google
23,250
Trustpilot
16,673
Clean and filter
Drop empty text
-9,352 Google
Parse datetime
extract temporal features
Negative filter
5,931 reviews
4,137 at 335 shared
locations
Multi-method analysis
BERTopic
4 lenses
full · top-30 · anger · LLM
Gensim LDA



Key Findings

Two Platforms, Two Complaint Cultures

The most significant structural finding is that Google and Trustpilot capture fundamentally different complaint types. Google reviews focus on the **physical gym experience**: equipment, cleanliness, overcrowding, music volume, and parking. Trustpilot reviews focus on the **business experience**: membership fees, cancellation difficulties, billing errors, and customer service. The word "membership" ranks #3 on Trustpilot negative reviews but does not appear in Google's top 10. Platform-specific bigrams confirm this: Google surfaces "free weights" and "peak times" while Trustpilot surfaces "joining fee" and "direct debit."



Running BERTopic separately on each platform produced 11 topics each, but with distinct compositions. Google found locker theft (51 reviews), loud music (43), and overcrowding (37) — all physical issues. Trustpilot found joining fee complaints (231 reviews), PIN/app problems (164), and customer service failures (38) — all business process issues.

Complaint Topics and Their Specificity

BERTopic identified 10 distinct complaint clusters from the merged negative reviews. Beyond the expected general complaints (1,443 documents), the model isolated highly specific, actionable issues: parking fines of exactly £85 (124 reviews), a 4-hour class cancellation policy causing frustration (115 reviews), locker theft and broken locks (112 reviews), and water machines broken for weeks without management action (38 reviews). This granularity is where BERTopic excels over LDA, which tends to merge related themes.

LDA's 10-topic model provided a useful complementary perspective. It automatically separated Danish (Topic 5) and German (Topic 7) language clusters — reviews from PureGym's Danish and Swiss

operations. It also produced a clear billing cluster (Topic 9: "membership, cancel, joining fee, payment") that aligned with Trustpilot's platform-specific topics.

Two further BERTopic runs sharpened the picture. The **anger-filtered run** (2,537 reviews, 24.8% outliers vs the full run's 35.6%) narrowed the primary anger drivers to membership cancellation, rude staff, and equipment failures — confirming the same themes the full run found but at higher resolution, since the anger filter strips out the more diffuse complaints. The **LLM-driven run**, where Qwen first extracted natural-language topic phrases per review and BERTopic then meta-clustered the 5,999 phrase outputs, produced substantively different clusters dominated by intent-bearing phrases like "personal turnover" and "rude staff feedback" — capturing customer meaning in a way bag-of-words BERTopic cannot.

Emotion Analysis Reveals Urgency Layers

Classifying all negative reviews by emotion showed that **44% express anger** (2,815 reviews), split evenly between platforms. Sadness accounts for 23%, and fear for 6%. The temporal dimension adds depth: anger and sadness peak at 8PM (post-gym frustration), but **fear peaks at 1AM** — late-night safety concerns at 24/7 gyms. Surprise reviews are the longest (median 58 words vs 34 for anger), suggesting unexpected experiences prompt more detailed accounts.

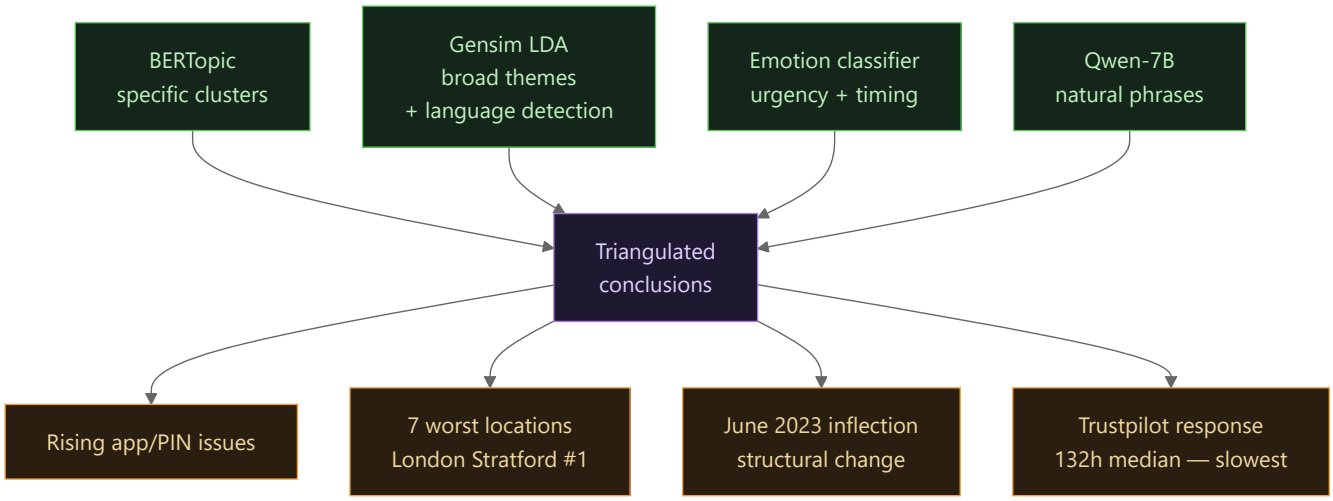
Star rating further differentiates emotions: 1-star reviews are 48% anger, while 2-star reviews shift toward sadness (25%) — the difference between a furious customer and a disappointed one, each requiring a different response strategy.

Temporal Trends: Things Are Getting Worse

Dynamic topic modelling revealed that **every complaint category is rising**. General complaints grew from 45 to 530 mentions across the analysis period. App and PIN problems — virtually nonexistent at the start — grew to 76 mentions, indicating a new systemic issue likely tied to a software update. Negative reviews tripled from June 2023 onwards, with April 2024 as the worst month (335 negative Google reviews). This trajectory suggests a structural change in operations, pricing, or staffing rather than isolated incidents.

Location Hotspots

Seven locations appear in both platforms' top 20 worst lists, confirming them as genuinely problematic rather than platform-specific anomalies. London Stratford leads with 81 combined negative reviews. The top-30 wordcloud sharpened against the full-dataset wordcloud — broad complaint terms gave way to location-specific vocabulary (`mould` , `closed` , instructor names) — confirming that the focusing effect happens at the wordcloud level too, not just in BERTopic. Running BERTopic on just the top 30 locations (37.1% outliers vs 35.6% for the full dataset) acted as a different lens, surfacing location-specific issues invisible at full scale: mould in showers, individual gym closures despite 24/7 advertising, and instructor-specific class complaints.



Conclusion

The power of this analysis lies in triangulation. BERTopic excels at specific, actionable complaint isolation. LDA provides interpretable broad themes and detects multilingual segments. Emotion classification adds urgency that pure topic modelling misses. Qwen-7B generates human-readable topic phrases that capture nuance — "charged after cancelling" conveys intent that no bag-of-words model can detect. Aspect-based sentiment analysis revealed cleanliness and safety with the highest negative sentiment rates (98%); parking was the least universally discussed aspect.

For PureGym, the immediate priorities are clear: address the rising app/PIN issues, improve response times for angry Trustpilot complaints (currently 132 hours median — the slowest, when they should be fastest), investigate the June 2023 inflection point, and focus operational improvements on the seven consistently worst locations. Each method pointed to the same conclusion from a different angle — which is precisely what makes multi-method NLP analysis valuable.

Appendix — Method comparison, churn signal, and LLM triangulation

Method-by-method comparison.

Method	Input	Outliers	What it surfaces
BERTopic — full	4,137 cross-platform negatives	35.6%	Specific clusters: parking £85, locker theft, app/PIN
BERTopic — top-30 worst	3,690 reviews from worst sites	37.1%	Location-specific: mould, 24/7 closures, instructor classes
BERTopic — anger-only	2,537 anger-classified reviews	24.8%	Most actionable: membership, rude staff, equipment
BERTopic — LLM-driven	5,999 Qwen-extracted phrases	9.8%	Natural clusters: "personal turnover", "rude staff feedback"
Gensim LDA	full negative reviews	n/a	10 broad themes (coherence 0.449); Danish + German detected

Why Qwen. Falcon-7B's rubric prompts no longer reproduce under post-update weights (model-version drift). Qwen-2.5-7B runs the same prompts deterministically on A100; the Sonnet-Qwen agreement below confirms operational equivalence.

Churn signal. The Twitter-trained emotion classifier labels ~1,486 British-understated 1-2 star reviews as "joy". A merged `churn_signal = emotion ∈ {anger, sadness} ∨ stars ≤ 2` recovers those OOD cases: on the 30,617-review full population it captures 20.9% as churn-risk versus 14.2% for emotion alone — 2,032 additional reviews that emotion-only misses.

Gold-vs-workhorse LLM comparison (30 held-out reviews). Claude Sonnet 4.6 produced gold-standard labels under a system prompt grounded in Perplexity Sonar Deep Research on PureGym's PE-ownership economics (40% first-year churn, 27% mature-site ROIC, £600M Leonard Green acquisition EV). Qwen 2.5-7B zero-shot and Qwen 2.5-7B 10-shot (with ten Sonnet-derived few-shot examples) were benchmarked against that gold. Operational-lever agreement rose from **60.0%** → **73.3%**, churn-risk agreement from **53.3%** → **70.0%**, primary-topic token Jaccard from **0.124** → **0.166**. Coaching a small open model with frontier-model examples closes most of the gap at zero marginal compute cost.

Cost. $1.10 sunk Perplexity Sonar Deep Research (2026-04-11) + $0.148 new Claude Sonnet 4.6 (40 calls: 10 iteration + 30 gold eval) + $0 Qwen on Colab Pro+ = $1.25 total, of which $0.148 is incremental. Full code and artefacts in `basic/basic_notebook_appendix.ipynb` at commit `81cede1`.