Applying NLP for topic modelling in a real-life context

Contents

Problem Statement

PureGym receives tens of thousands of customer reviews across platforms like Google and Trustpilot. While many are positive, a significant portion express dissatisfaction—often without clear patterns. This project uses Natural Language Processing (NLP) to identify key drivers of negative feedback. By uncovering recurring topics and emotional signals in ~40,000 reviews, we aim to help PureGym address member concerns and improve service delivery.

Summary of Approach

We combined reviews from Google (23,250) and Trustpilot (16,673), applied text cleaning and tokenisation, and focused on low-rated entries. BERTopic was first used to extract themes from negative reviews. We then applied a BERT-based emotion classifier to isolate “anger”-dominated cases for separate modelling. LLM-based topic extraction failed, so we used ChatGPT to simulate structured topic output, which was clustered using BERTopic. Finally, we compared results against traditional LDA using Gensim.

Initial Topic Modelling with BERTopic

We applied BERTopic to the full set of negative reviews, clustering texts by semantic similarity and extracting representative keywords. This process produced 246 topics.

Figure 1 shows the intertopic distance map, where each bubble represents a topic and its size reflects its prevalence. Many bubbles overlap, making the visual difficult to interpret. Filtering to highlight only the most dominant topics could improve clarity.

Figure 2 displays the top keywords for selected topics. Words like “clean”, “tidy”, and “equipment” appear across multiple clusters, pointing to recurring concerns about cleanliness and facilities. Other topics reflect issues such as staff behaviour, pricing, and membership cancellations. These results provided an initial overview of customer dissatisfaction themes.

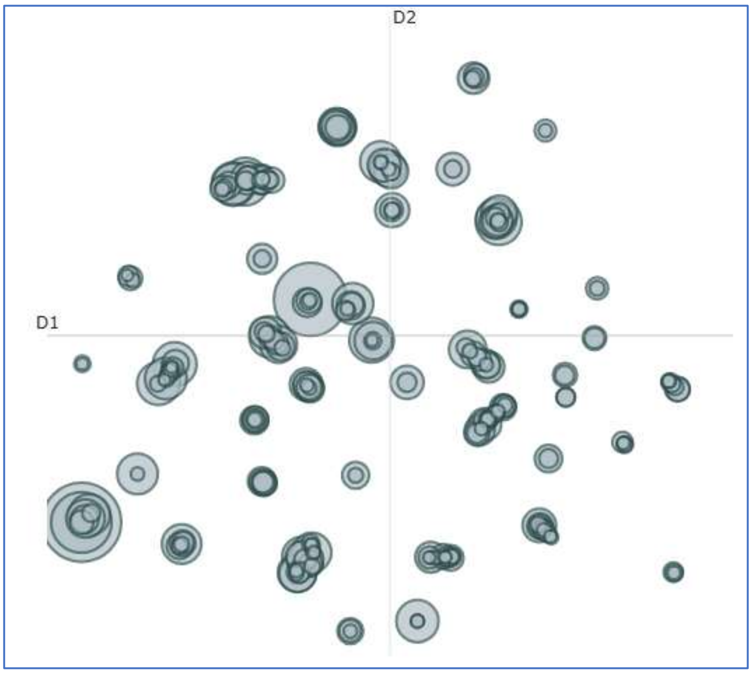

Figure 1: Intertopic Distance Map

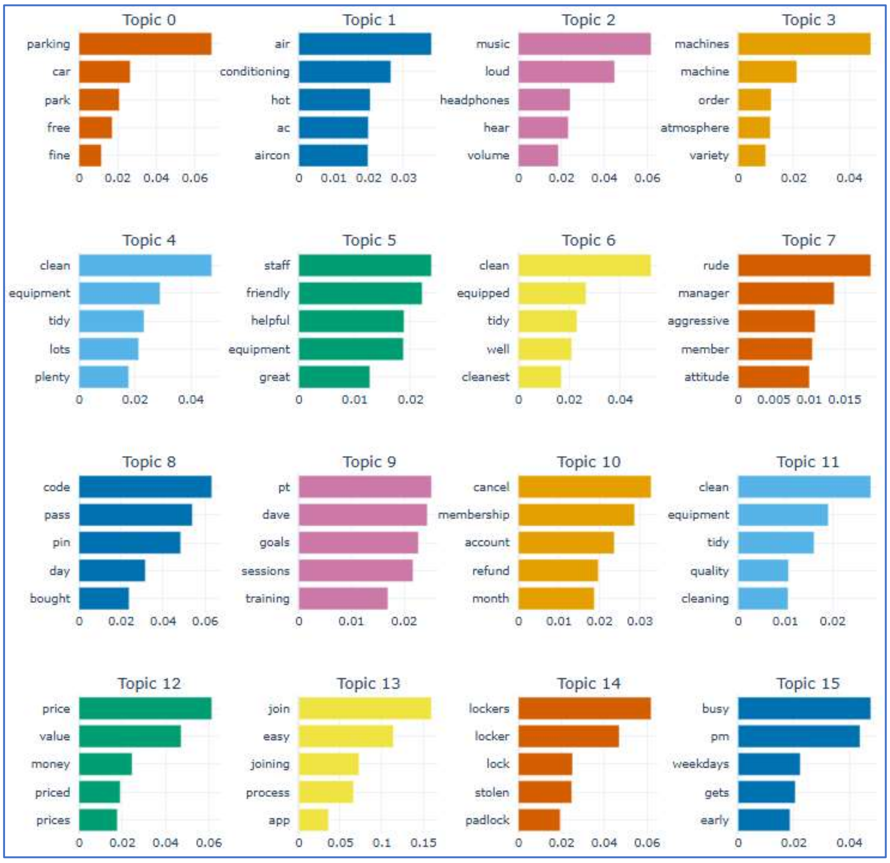

Figure 2: Frequency Bar Chart of Top Topic Keywords

Emotion Detection and Anger-Focused Analysis

We used a BERT-based emotion classifier to assign each review a dominant emotion. As shown in **Figure 3**, anger was the most common emotion among low-rated reviews across both platforms.

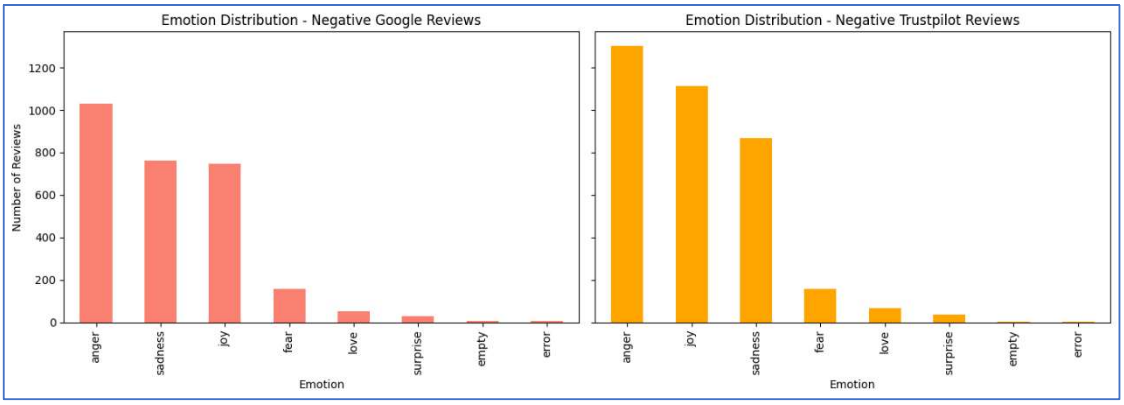

Figure 3: Distribution of dominant emotions in low-rated reviews

Filtering for anger-dominated reviews, we applied BERTopic separately to Google and Trustpilot subsets. **Figure 4** shows the top keywords for the resulting clusters, while **Figure 5** presents their spatial layout.

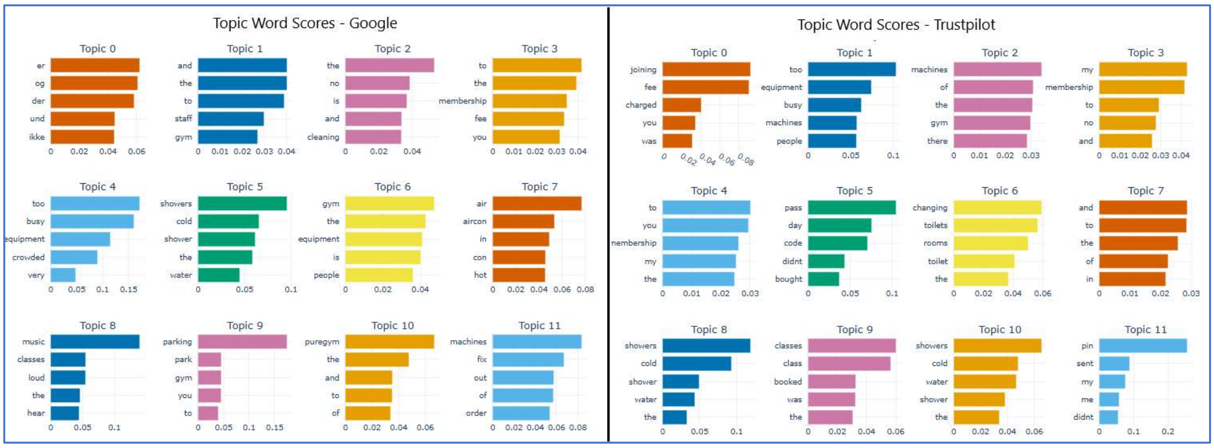

Figure 4: Top keywords for dominant BERTopic topics in anger-dominated reviews (Google and Trustpilot)

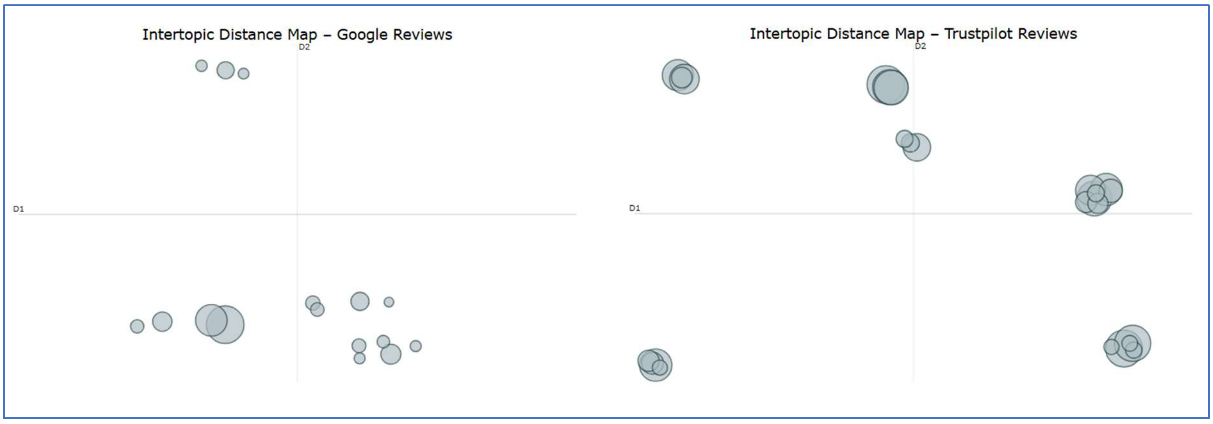

Figure 5: Intertopic distance maps for anger-dominated reviews in Google (left) and Trustpilot (right)

In the Google reviews, themes included overcrowding (Topic 4), showers (Topic 5), cleanliness (Topics 2), ventilation (Topic 7), noise (Topic 8), parking (Topic 9), and broken equipment (Topic 11). Some noise appeared in Topics 0 and 1.

Trustpilot topics showed similar themes but were more clearly defined, including billing (Topic 0), hygiene (Topic 6), class booking (Topic 9), and PIN access (Topic 11). These anger-focused models revealed consistent pain points with sharper clarity.

LLM Experiments and Clustered Insights

We initially attempted to extract topics using hosted LLMs (Falcon, Phi, FLAN-T5), but these failed due to technical errors or unstructured outputs. To continue, we used ChatGPT to simulate topic extraction, prompting it to return three numbered topics per review. This produced over 1,000 topic phrases.

We applied BERTopic to this set of phrases, treating them as a new corpus. The resulting clusters provided a higher-level summary of recurring themes across reviews.

Table 1 shows a sample of phrases and their cluster assignments. For example, "Gym equipment" grouped into Cluster 1, while "Staff behaviour" appeared in Cluster 0, demonstrating that the model captured coherent categories even from short inputs.

Topic Phrase	BERTopic Cluster
Gym equipment	1
Cleanliness	2
Overcrowding	3
Temperature control	5
Opening hours	10
Staff behaviour	0
Staff behaviour	0
Temperature control	5
Gym equipment	1
Gym equipment	1

Table 1: Sample topic phrases and their BERTopic cluster assignments (from ChatGPT-extracted inputs)

Figure 6 shows the frequency distribution of clusters, highlighting which themes were most prominent across the review set.

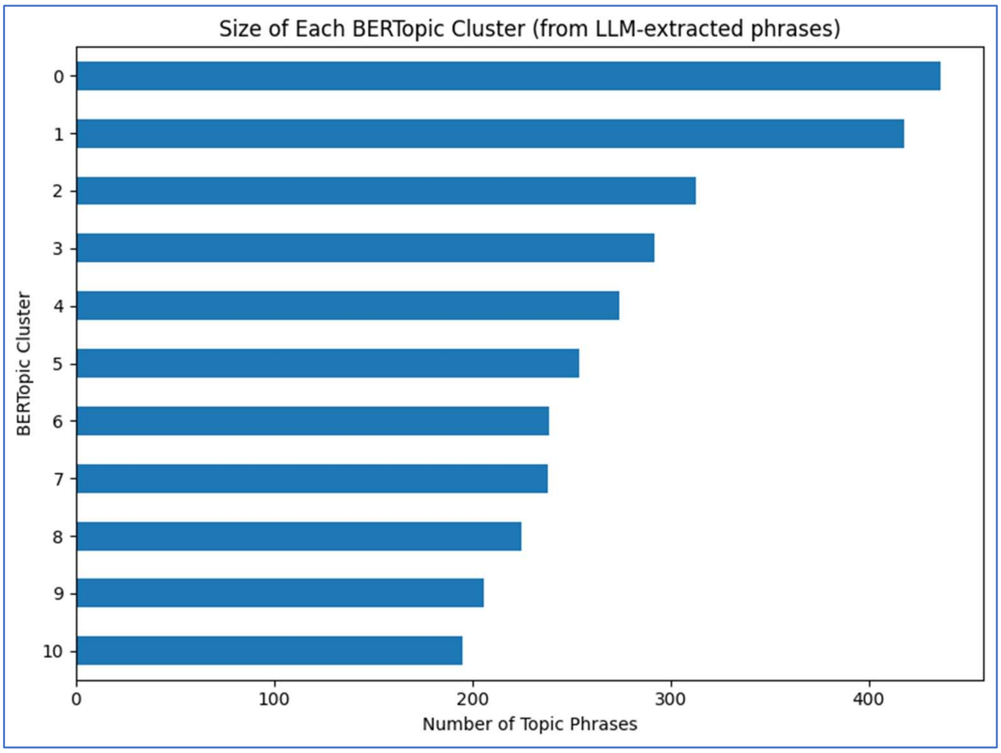

Figure 6: Frequency distribution of BERTopic clusters derived from ChatGPT-extracted topic phrases

Comparison with Gensim LDA

We applied Latent Dirichlet Allocation (LDA) using Gensim as a benchmark against BERTopic and the LLM-based approach. The model was trained on combined negative reviews from Google and Trustpilot, using the same preprocessing pipeline.

Neutral terms like “gym” and “equipment” were excluded, and the model was configured to extract ten topics. Figure 7 shows the intertopic distance map and the 30 most salient words.

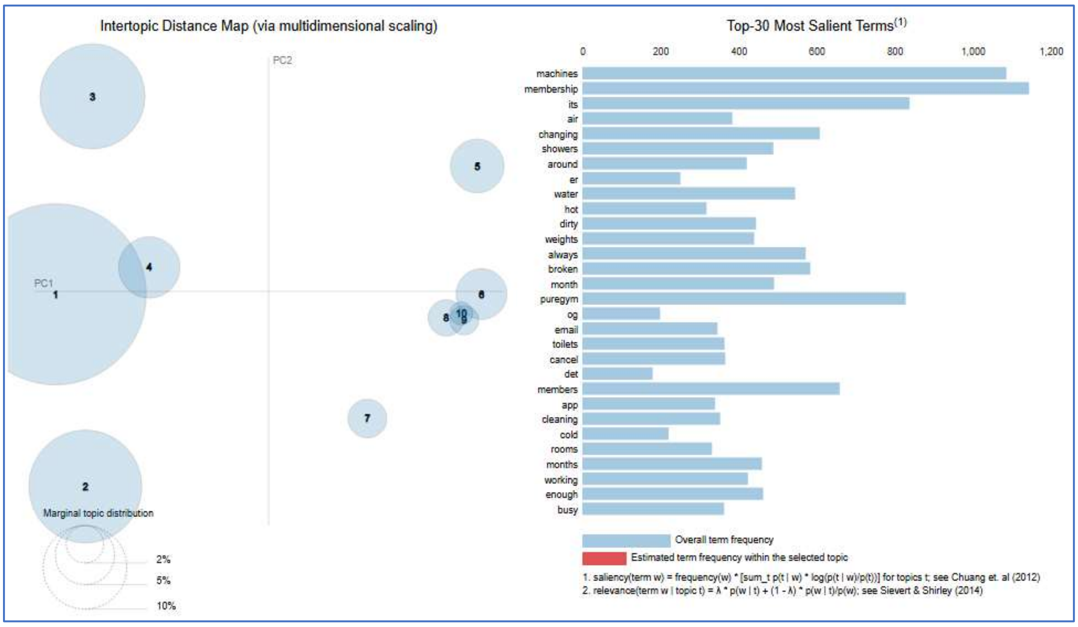

Figure 7: LDA visualisation showing topic distances (left) and most salient terms (right)

The LDA model highlighted operational concerns such as “machines”, “membership”, “changing”, “showers”, and “air”. Some noise remained (e.g., “its”, “og”), but the results were more balanced and diverse. The distance map showed clear topic separation, suggesting the model captured distinct complaint themes at a broader scale than BERTopic.

Conclusion

This project compared multiple topic modelling techniques to analyse negative customer reviews from Google and Trustpilot. The aim was to identify recurring themes of dissatisfaction and assess the interpretability of different methods.

BERTopic, when applied to all negative reviews, produced 246 clusters but suffered from visual overcrowding. Filtering by emotional tone improved clarity. After applying a BERT-based classifier, we isolated anger-dominated reviews, revealing clearer topics such as broken equipment, poor hygiene, and staff behaviour—especially in the Trustpilot data.

To simulate LLM-based topic extraction, we prompted ChatGPT to return three topics per review. Clustering these phrases with BERTopic produced interpretable groupings like “Gym equipment” and “Staff behaviour,” showing the method’s utility even on brief inputs.

Gensim LDA offered a traditional alternative, returning ten topics with salient terms such as “membership”, “machines”, and “showers”. Despite some noise, it delivered a more balanced view of the full corpus.

Each method brought different strengths: BERTopic excelled on focused or structured inputs, while LDA performed well on broader data. Together, they revealed consistent operational issues driving customer dissatisfaction.