

Applying NLP for Topic Modelling and Actionable Insight Generation



UNIVERSITY OF
CAMBRIDGE

Anastasia Alexandra Krivenkovskaya

Institute of Continuing Education

University of Cambridge

This academic report is submitted for the course

Applying advanced data science techniques P1_2025

CAM_DS_301

August 2025

Table of Contents

INTRODUCTION	3
DATA OVERVIEW AND PREPROCESSING	3
BERTOPIC MODELLING.....	5
EMOTION ANALYSIS	6
FALCON-7B-INSTRUCT AND INSIGHT GENERATION	7
TOPIC MODELLING WITH LDA (GENSIM).....	9
DISCUSSION AND INTERPRETATION.....	10
FINAL INSIGHTS AND CONCLUSION.....	11

Introduction

This report presents a comprehensive application of Natural Language Processing (NLP) techniques to identify and analyze major concerns expressed in customer reviews of gym locations. The goal is to extract structured insights from unstructured textual feedback using emotion classification, topic modelling, and large language model generation. Two main sources of reviews were considered: Google and Trustpilot.

The project employs a range of modern tools including BERTopic for unsupervised topic modelling, Falcon-7b-Instruct for generating high-level insights, and visualization techniques to facilitate interpretation. The findings aim to assist gym operators in addressing recurring customer issues through data-driven decision-making.

Data Overview and Preprocessing

The dataset comprises negative reviews collected from Google and Trustpilot. Reviews were filtered to retain only those written in English using the langdetect package. After cleaning, a total of 2784 anger-related reviews were analyzed: 1193 from Google and 1591 from Trustpilot.

A bar chart was generated to visualize the most frequent terms across the top 30 locations (Figure 1). Additionally, a word cloud was created to represent the most common words (Figure 2), with terms such as *equipment*, *staff*, *machines*, and *membership* appearing prominently.

BERTopic Modelling

Two BERTopic models were trained:

- One on all negative reviews (Figure 3),
- Another focused on reviews from the top 30 most-reviewed locations (Figure 4).

Each BERTopic model produced a list of clustered topics with corresponding word distributions and representative documents. For example, the most prominent topics across reviews included *air conditioning*, *dirty changing rooms*, *equipment issues*, and *membership fees* (Figure 3).

To further refine analysis, BERTopic was also applied to reviews classified as **anger**. Clusters identified in this subset emphasized *rude staff*, *unhygienic showers*, and *lack of enforcement* (Figure 6).

The following comments were derived from these analyses:

- Clusters based on emotional filtering were more specific and actionable.
- Emotion-based BERTopic results offered greater clarity around operational pain points.

BERTopic was applied to both the general set of negative reviews and the subset corresponding to the top 30 most negatively reviewed locations. The model was configured with a custom CountVectorizer using bi-grams, and a minimum topic size of 10 to capture meaningful clusters.

Topic	Count	Name	Representation	Representative_Docs
0	221	0_air_conditioning_the_hot	[air, conditioning, the, hot, is, aircon, con,...	[Normally, I never leave a negative review as ...
1	164	1_my_membership_fee_to	[my, membership, fee, to, month, me, joining, ...	[I choose this gym as it seemed to offer a lot...
2	129	2_toilets_toilet_dirty_changing	[toilets, toilet, dirty, changing, smell, no, ...	[Changing rooms and especially toilets and sho...
3	124	3_parking_car_park_free	[parking, car, park, free, fine, you, to, gym,...	[Such a shame to have to write the review beca...
4	121	4_class_classes_was_to	[class, classes, was, to, the, booked, instruc...	[Today was my first day and really let down. R...
5	112	5_lockers_locker_and_the	[lockers, locker, and, the, my, room, to, are,...	[This gym is lucky to even get one star. Dam...
6	95	6_closed_open_247_gym	[closed, open, 247, gym, on, hours, to, is, op...	[I have been a member of this gym for 6 years ...
7	94	7_showers_cold_shower_water	[showers, cold, shower, water, the, to, temper...	[When I first joined PureGym the showers were ...
8	92	8_is_the_price_for	[is, the, price, for, gym, of, to, puregym, an...	[Pros: - It's a new gym, so everything is new,...
9	76	9_her_she_me_was	[her, she, me, was, to, and, he, this, rude, m...	[Avoid this gym if you want to exercise in a f...

Figure 3: BERTopic Output for All Locations

Topic	Count	Name	Representation	Representative_Docs
-1	213	-1_gym_people_equipment_just	[gym, people, equipment, just, weights, machin...	[I have been a member of this gym for 6 years ...
0	95	0_gym_staff_manager_rude	[gym, staff, manager, rude, member, puregym, d...	[Avoid this gym if you want to exercise in a f...
1	63	1_showers_shower_water_fix	[showers, shower, water, fix, cold, dirty, gym...	[Stark comparison - Pure Gym St Georges Leice...
2	59	2_gym_toilets_cleaning_dirty	[gym, toilets, cleaning, dirty, disgusting, cl...	[So I was recently asked by Pure Gym to be sub...
3	55	3_air_gym_ac_aircon	[air, gym, ac, aircon, conditioning, air condi...	[I used to love this gym, having been a member...
4	50	4_pin_pass_day pass_email	[pin, pass, day pass, email, work, didnt, day,...	[Only giving 1 star as there is no option for ...
5	47	5_broken_busy_equipment_overcrowded	[broken, busy, equipment, overcrowded, machine...	[Everything always broken, Horrendously overcr...
6	41	6_parking_car_free_85	[parking, car, free, 85, free parking, fines, ...	[As several others have mentioned, members are...
7	33	7_lockers_locker_room_locker room	[lockers, locker, room, locker room, gym, take...	[For a newly built gym the management should b...
8	31	8_gym_equipment_youre_people	[gym, equipment, youre, people, machines, spac...	[Zero enforcement of any rules.\n\nEquipment i...

Figure 4: BERTopic Output for Top 30 Locations

Emotion Analysis

Emotion detection was performed to isolate reviews most likely to signal dissatisfaction. Using the [bhadresh-savani/bert-base-uncased-emotion](#) model from Hugging F, the top emotion identified in both datasets was **anger**, followed by **sadness** and **joy** (Figure 5).

This emotional filtering proved crucial for subsequent topic extraction, as it narrowed the scope of modelling to the most critical subset of feedback.

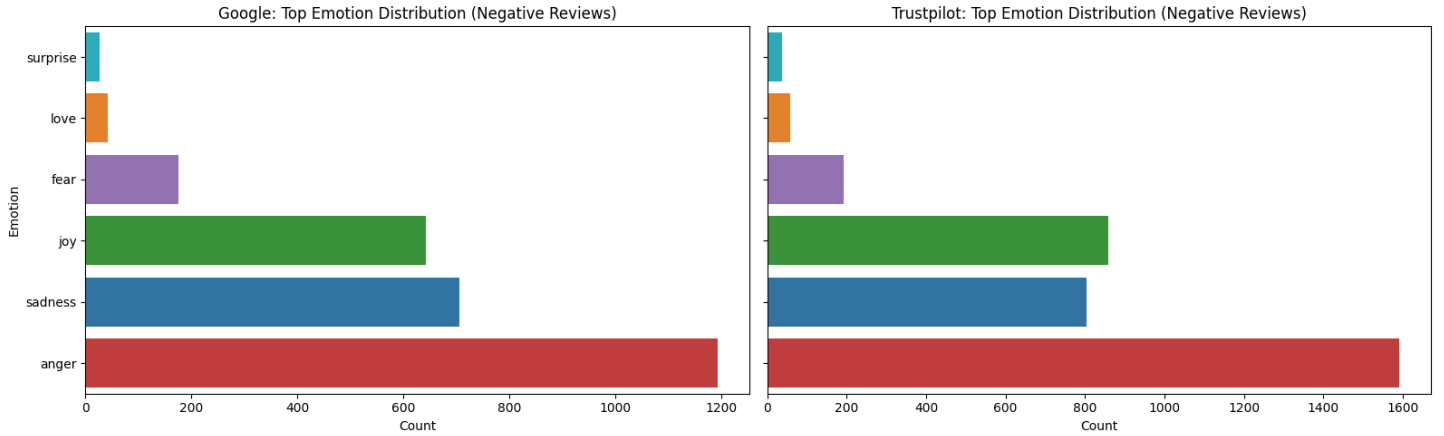


Figure 5: Emotion Distribution in Negative Reviews

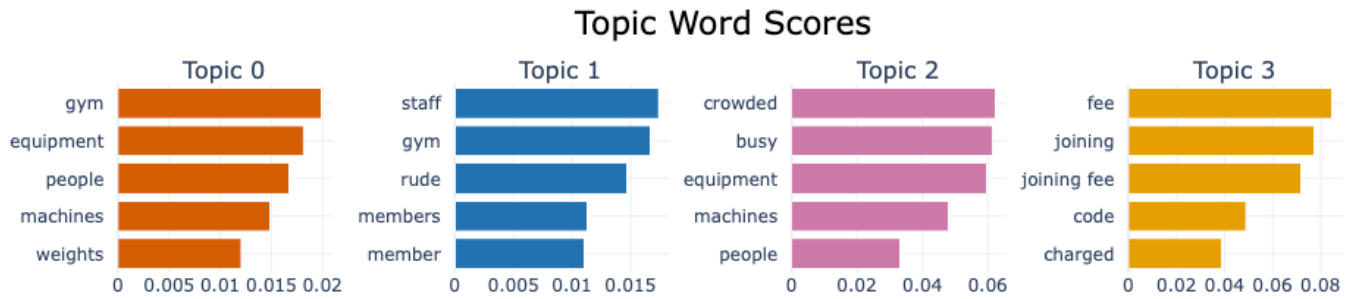


Figure 6: BERTopic Clusters for Angry Reviews

Falcon-7b-Instruct and Insight Generation

To extract high-level suggestions for each topic, the Falcon-7b-Instruct model was employed in two stages:

1. **Topic Extraction:** Each review was prefixed with the instruction:

"In the following customer review, pick out the main 3 topics. Return them in a numbered list format, with each one on a new line."

The output was parsed to extract discrete topics from each review.

2. **Insight Generation:** Using the extracted topics, a second prompt was used:

"For the following text topics obtained from negative customer reviews, can you give some actionable insights that would help this gym company?"

The outputs were consolidated into a suggestions table (Figure 7), which included topic-suggestion pairs. Example insights include:

- *Improve transparency in cancellation policies.*
- *Increase investment in staff training for better customer interaction.*
- *Upgrade outdated equipment and offer maintenance incentives.*

These generated insights support direct action and can complement human-led strategic planning.

Topic	Suggested Action
- Amenities offered:	1. Variety of equipment. 2. Cleanliness. 1. Offer different types of equipment to cater to customers with various fitness goals and preferences. 2. Implement stringent cleaning protocols to maintain a safe and hygienic environment for members.
- Cleanliness of the gym:	1. Implement a strict cleaning protocol to ensure surfaces, equipment, and personal belongings are regularly sanitized. 2. Train staff to maintain a high level of cleanliness, and create a visual inspection process to identify and address any uncleanness promptly. 3. Offer incentives or rewards to motivate members to maintain a clean and tidy environment. - Gym equipment: 1. Perform routine maintenance on equipment to prevent breakdowns and ensure optimal performance. 2. Offer regular cleaning and inspection services to keep equipment hygienically maintained. 3. Invest in high-quality, durable equipment to minimize mechanical issues and increase longevity. - Staff training and customer service: 1. Develop and implement a comprehensive staff training program to ensure knowledge and competency in customer service, cleanliness, and proper usage of equipment. 2. Train staff to actively engage with members, provide personalized attention, and foster a welcoming environment. 3. Implement a customer feedback system to track member satisfaction and identify areas of improvement. - Marketing and branding: 1. Create a strong online presence with a visually appealing website, social media pages, and relevant content. 2. Implement a targeted marketing strategy to attract new members, including leveraging local community events and influencers. 3. Sponsor local events, athletes, or organizations to increase brand visibility and promote the gym's services.
- I'm going to give it a pass.	1. Lack of attention to customer comfort and experience: Gyms should ensure that their facilities are well-maintained, clean, and adequately equipped to provide a comfortable and welcoming environment for their patrons. Regularly updating the equipment and providing fresh, comfortable attire can go a long way in creating a positive customer experience. 2. Unfriendly staff: Gym staff should be well-trained in customer service and ideally should be friendly, welcoming, and knowledgeable. Interactions with customers should be prompt and personalized to create a strong customer relationship. 3. Poor communication: Gyms should ensure that their staff communicates effectively with their customers, addressing any issues that arise and keeping them updated on any changes or updates at the gym.
- Over-played music	1. Implement sound masking to reduce noise levels. 2. Encourage customers to purchase noise-cancelling headphones. 3. Provide earplugs to reduce the volume. - Poor equipment maintenance 1. Regularly schedule maintenance checks to ensure equipment is functioning properly. 2. Offer equipment replacement or repair services. 3. Implement a rewards program for members who regularly maintain their equipment. - Inefficient staff training 1. Invest in proper staff training to ensure employees can effectively handle customer inquiries and complaints. 2. Implement customer feedback surveys to identify areas of improvement. 3. Offer continuing education opportunities for staff to keep up with industry trends. - Lack of cleanliness 1. Implement regular cleaning schedules for equipment and common areas. 2. Offer incentives for members to maintain their personal hygiene. 3. Create a visual cue system to encourage members to clean their equipment after use.
- Overall rating:	2 stars 1. Improve customer service: Responding to negative reviews and addressing customers' concerns in a timely manner can go a long way in improving customer satisfaction, ultimately leading to better ratings. 2. Take action on feedback: Implementing the feedback from customers can help improve the gym's offerings, leading to increased loyalty and better ratings over time. 3. Enhance the gym experience: Offering additional amenities such as personal training, group classes, or additional equipment can help make the gym experience more appealing to customers, ultimately leading to higher ratings. 4. Communicate transparently: Providing updates or explanations to customers on any issues or concerns can help build trust and improve customer satisfaction, leading to better ratings. 5. Offer discounts or promotions: Offering limited-time deals or discounts to customers can help attract new customers and improve customer retention, ultimately leading to higher ratings.
- Poor customer service	1. Implementing a customer-focused approach to customer service, including hiring and training staff with exceptional customer service skills and providing regular customer service training. 2. Offering flexible class schedules and personalized training plans to cater to individual needs and preferences, which could lead to increased customer satisfaction and loyalty. 3. Creating a user-friendly online platform for customers to easily access and manage their accounts, schedules, and class reservations. 4. Offering discounts and promotions to attract new customers, as well as providing incentives for existing customers to maintain their loyalty. 5. Investing in high-quality equipment and amenities to enhance the overall customer experience and differentiate the gym from competitors.

Figure 7: Generated Suggestions from Falcon-7b-Instruct

Topic Modelling with LDA (Gensim)

LDA was applied using Gensim on tokenised negative reviews. Ten topics were specified. Preprocessing involved lemmatisation, stopword removal, and dictionary creation. The output was visualised via pyLDavis:

- **Figure 8:** LDA Intertopic Distance Map
- **Figure 9:** Top-30 most salient terms

The topics include frequent terms like "gym", "equipment", "shower", "membership", and "staff". Unlike BERTopic, LDA does not produce labelled topics automatically and requires interpretation based on top terms.

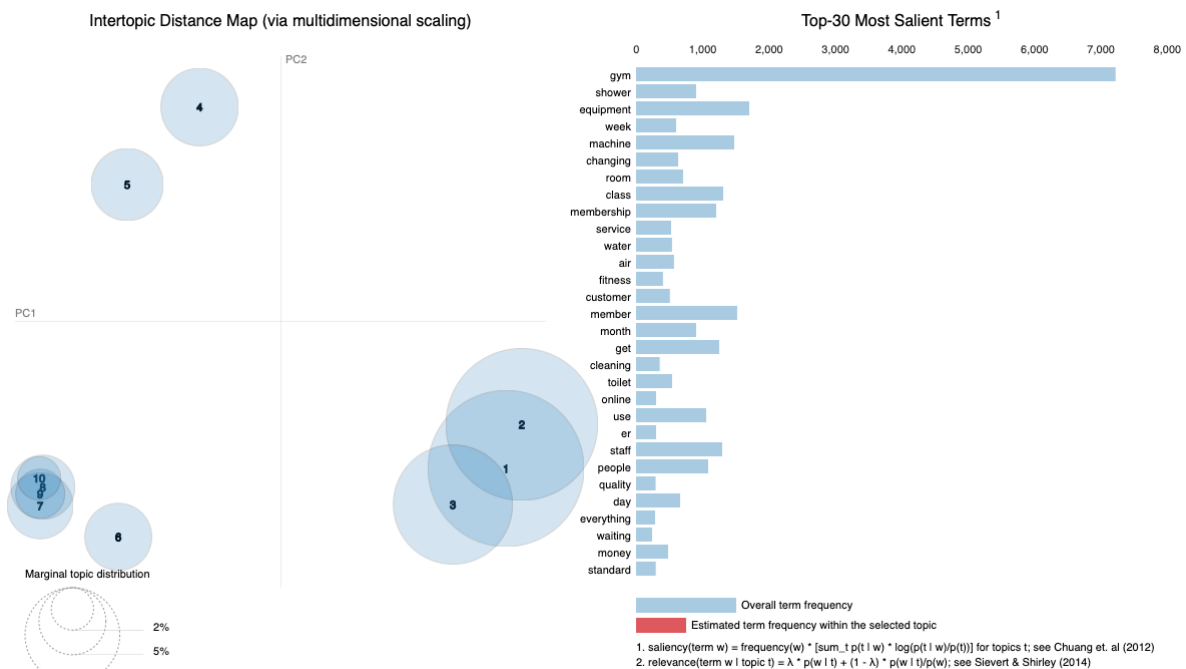


Figure 8 & 9: LDA Intertopic Distance Map and Top-30 most salient terms

Discussion and Interpretation

Comparison of Techniques:

- **BERTopic** produced interpretable and labelled clusters. It leveraged contextual embeddings, resulting in granular and actionable themes (e.g. cold showers, parking fines).
- **LDA** was less intuitive in terms of labelling. Topics are numerically indexed and require manual inspection. However, the word distributions largely validate key issues surfaced by BERTopic.
- **Emotion detection** helped prioritise emotionally charged reviews. This refined BERTopic clustering and insight generation.

Interpretation of Visualisations:

- The **intertopic distance map** (LDA) showed reasonable separation, with some overlap indicating semantically close topics.
- The **salient term bar chart** confirmed the recurrence of gym-related complaints: equipment, showers, staff, and cleanliness.

Insight Generation:

- **Figure 7:** Suggestions generated using Falcon-7b based on extracted topics. Topics such as *air conditioning*, *customer service*, *membership suspension*, and *poor hygiene* were translated into targeted improvement ideas.

These insights can help gym managers address critical issues such as infrastructure upgrades, staff retraining, and transparent membership policies.

Final Insights and Conclusion

Through a combination of topic modelling and emotion analysis, several patterns have emerged:

- **Core themes** include: poor customer service, equipment maintenance, cleanliness, overcrowding, and inflexible pricing.
- **BERTopic** enabled human-readable labelling, while **LDA** validated these clusters statistically.
- **Emotion analysis** highlighted customer anger and sadness, validating the severity of complaints.
- **Generated suggestions** translated customer pain points into business actionables.

In summary, the project demonstrates how modern NLP models can extract meaningful and business-relevant themes from large textual datasets. Each method contributed uniquely to the interpretation, and their integration offered a fuller picture of customer dissatisfaction.