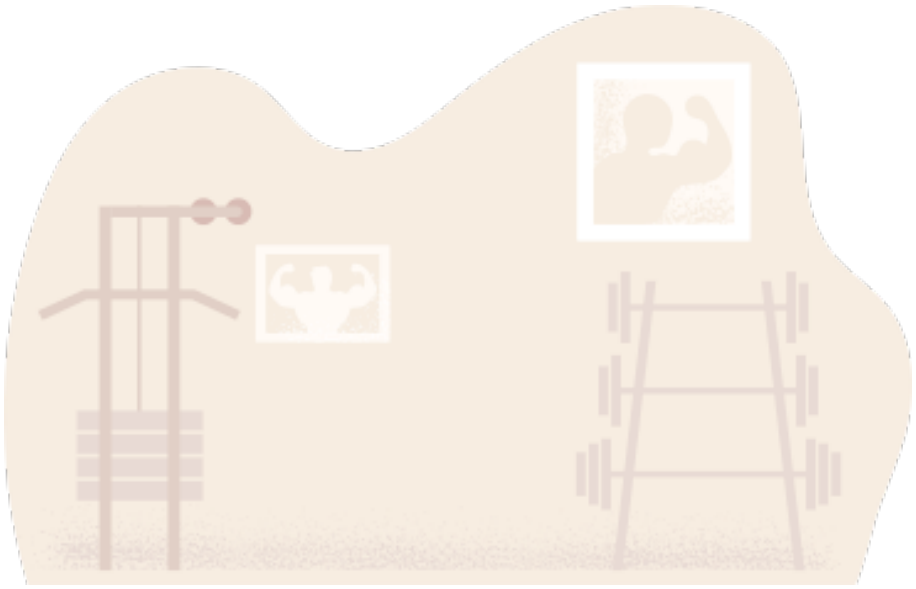
Navigating the Gym Landscape: Analyzing Customer Reviews in Danish Facilities

CHRISANNA K. CORNISH CCOR@ITU.DK
CHRISTIAN M. HANSEN CHMH@ITU.DK
CONSTANTIN-BOGDAN CRACIUN COCR@ITU.DK
GINO F. FAZZI GIFA@ITU.DK
VERON HOXHA VEHO@ITU.DK



1ST SEMESTER, AUTUMN 2023
DATA SCIENCE MSC
KSDWWVD1KU

IT UNIVERSITY OF COPENHAGEN

Navigating the Gym Landscape: Analyzing Customer Reviews in Danish Facilities

Chrisanna K. Cornish
IT University of Copenhagen
ccor@itu.dk

Christian M. Hansen
IT University of Copenhagen
chmh@itu.dk

Constantin-Bogdan Craciun
IT University of Copenhagen
cocr@itu.dk

Gino F. Fazzi
IT University of Copenhagen
gifa@itu.dk

Veron Hoxha
IT University of Copenhagen
veho@itu.dk

Abstract

Customer reviews play a pivotal role in shaping consumer decisions across a spectrum of industries, offering indispensable insights into the quality, dependability, and trustworthiness of products, services, and companies. Online review platforms act as democratized audits, fostering the exchange of knowledge and experiences among customers. In this study, we turn the focus towards a specific industry: fitness facilities in major Danish urban areas. Beyond exploring rating variations across fitness brands, our research extends its scope to delve into significant differences across some common review topics. We present a comprehensive framework designed to methodically collect, process, analyze, and visualize insights derived from customer reviews within Danish fitness facilities. Despite the inherent constraints imposed by sample size limitations, our research unearths noteworthy variations in customer satisfaction across companies and review platforms. Moreover, the study identifies different biases and limitations in the data and process pipeline, providing a foundation for heightened awareness, and lay the groundwork for future improvements in the process of understanding customer experiences in the fitness industry.

The necessary code and data to replicate this report is available on: Data Wild West - GitHub Repo.

1 Introduction

Customer reviews have become a major topic of research for brands of all industries, and a significant driving force influencing consumers' decisions. Acting as a form of social validation, these reviews offer prospective customers invaluable perspectives on a product, service and even a company's quality, dependability, and trustworthiness. Companies are aware that favorable feedback has the power to elevate a brand's standing, attract new customers, and cultivate trust. Previous research shows that positive reviews have a significant effect on consumer purchase behavior (Mo et al., 2015). Conversely, unfavorable reviews can wield a damaging effect, pinpointing areas for enhancement and potentially steering customers elsewhere.

In an age characterized by extensive consumerism, the internet facilitates the exploration of seemingly endless options, drowning customers in a sea of information. Consequently, review websites function as a democratized audit of brands, providing a platform for clients to exchange knowledge and share their experiences. As stated by (Karakaya and Ganim Barnes, 2010) "Companies can utilize the information provided by consumers on a variety of web sites ranging from blogs to rating and review sites to understand customer concerns and complaints so that they can take corrective actions".

In this context, we explore the world of reviews for a very specific industry: fitness facilities in the main Danish urban areas, mining customer reviews from two main sources, and classifying these reviews into buckets of interest ("topics"). More concretely, we want to know **"Do customer reviews sentiment differ significantly across fitness brands in Denmark?" In addition, we want to investigate if there is a significant difference across different review platforms.**

In Section 2 we address the main methods used to collect, clean and pre-process the data. In Section 3 we showcase our dataset in some potential downstream applications and provide statistical analysis on the review sentiment and rating. Finally, in Section 4, we delve into the potential biases present in our dataset, and disclose the primary limitations and cautions associated with its utilization.

2 Methods

2.1 Data collection

In the pursuit of comprehensive and accurate data, the collection process for this study hinged on a meticulous amalgamation of information from diverse sources. The primary supply of data comprised Google Maps and Trustpilot[1], two platforms renowned for their extensive coverage and user-generated content. Additionally, we enriched the dataset by extracting alternative facilities from the official repository of Copenhagen Municipality, via its website, serving as a cornerstone for local context, and a point of comparison for main brands. After locating these facilities, we retrieved the reviews associated to them.

To harness the wealth of review data offered by Google Maps, the study leveraged the Google Maps API, ensuring a systematic and standardized approach. Simultaneously, for other pertinent information residing on the web, a scraping methodology was employed to extract data from these sources. This judicious blend of proprietary API utilization and web scraping techniques laid the foundation for a robust and comprehensive dataset, fostering a nuanced understanding of the subject matter under investigation. Moreover, this approach not only facilitated the present study but also established a framework for the systematic extraction of a continuous stream of reviews from these sources, laying the groundwork for future research endeavors.

2.1.1 Google Maps API

The Google Maps API was a crucial tool in our data collection strategy, enabling us to gather detailed information about various establishments. Specifically, we focused on collecting data points such as reviews, authors, ratings, and precise geolocations for places of interest.

Key aspects of Google Maps API usage:

- **Service Endpoint/Base URL:** Our requests were directed to the Google Maps API service endpoint, the primary interface for accessing the data services provided by Google Maps[2].
- **Authentication Method:** We used an API key for authentication, a unique identifier necessary for authorizing data requests. This API key was obtained through registration with the Google Cloud Console.
- **Data Points Extracted:**
  - **Review:** Textual feedback from users.
  - **Author:** Names of the individuals who provided the reviews.
  - **Rating:** User ratings on a scale from 1 to 5.
  - **Coordinates of the Place:** Geographical locations represented by latitude and longitude.
- **Response Format and Parsing:** The API's responses, in JSON format, were parsed to extract relevant information. This involved systematic extraction of the required data points for our analysis, including the mentioned locations, review texts, author names, and ratings.

Despite utilizing the free version of the Google Maps API, which comes with certain limitations, it proved sufficient for our project requirements. It is noteworthy that our dataset from Trustpilot lacks geolocation information since the platform focus on brand-level reviews instead of specific businesses or facilities.

2.1.2 Trustpilot Reviews

Since the review-hosting website Trustpilot only provides paid API services for businesses, the requisite reviews were obtained through web scraping. Utilizing the BeautifulSoup library, the scraping process was directed toward specific enterprises, namely the prominent fitness brands PureGym, Sats, and Vesterbronx. For each brand, all available review pages were systematically collected, focusing on relevant HTML elements. The extracted information was then organized according to the fields outlined in Appendix B, Table 1. Furthermore, regular expressions played a crucial role in capturing the star ratings accompanying each review, adding a layer of precision to the data collection process.

2.1.3 Facilities retrieval from Copenhagen Municipality Webportal

The Copenhagen Municipality web portal serves as an extensive repository of public facilities and activities accessible to all citizens, offering an exhaustive selection of fitness-related facilities that could later on serve as a comparison to mainstream

[1] Google Maps is the widely know web mapping platform offered by Google; Trustpilot is a popular Danish review website which hosts reviews of businesses worldwide.

[2] *Google Maps Documentation* (Google Maps API Docs)

brands. Focusing specifically on the Copenhagen Municipality area, we identified an interactive map that aggregates diverse exercise and activity locations within the municipality[3]. A table, delineating the attributes of each location, can be found in the Appendix B (see Table 2). During data collection, a noteworthy challenge arose from the dynamic generation of the table, dependent on the activity categories displayed on the map. To overcome this, we employed the Selenium library to interact with the website through a webdriver. The full XPath of each relevant element was used to scrape the required fields. This presented its own difficulties, as some table entries were missing fields that a defined XPath would target, so this had to be accounted for.

2.1.4 Collection results

The final merged dataset comprises 3,586 reviews on fitness facilities from main cities in Denmark, including the review text, rating, author and enterprise[4]. A full description of the dataset attributes can be found in Appendix B (see Table 3).

2.2 Review Translations

Our dataset primarily consists of reviews from Danish establishments, a significant portion of which were originally written in Danish. Consequently, our dataset contains content in both English and Danish languages. While this bilingual aspect might not pose a problem in some scenarios, managing multilingual data can be challenging for various downstream tasks. Therefore, we opted to standardize the entire dataset by translating all Danish reviews into English. To accomplish the translation task, we chose to utilize an open-source translation model from the well-known NLP collaboration platform, **Hugging Face**[5]. This decision enabled us to seamlessly integrate the translation step into our data collection pipeline, leading to a more efficient and automated collection process. Upon visual inspection, the model appears to perform well; however, to accurately assess its performance, an additional round of manual annotations by native Danish speakers was necessary.

To measure the quality of the machine translation against the human translations, we rely on one of the metrics discussed in (Lee et al., 2023), Word Error Rate (WER). Word Error Rate is an untrained, lexical, word-based metric that is commonly used in the field, intuitive and easy to implement. We discuss the metric in detail in the next section.

2.2.1 WER Metric

The Word Error Rate (WER) serves as a widely used metric to assess the performance of speech recognition or machine translation systems, and it is calculated from the Levenshtein distance. Word error rate can then be computed as:

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct words, N is the number of words in the reference ($N = S + D + C$). The WER score has a lower bound of 0, indicating perfect match, but has no upper bound since N is the number of words in the reference text, and thus, the difference can be larger than the size of the reference text. In general, a lower WER indicates better performance, and values close to 1 indicate that every word in the reference is incorrect or missing in the system's output.

For our translations, we calculate the average WER score across reviews. The final average score is 0.388, suggesting that the machine translations do not differ considerably from the human translations, specially considering some understandable differences in the use of synonyms or verb conjugations.

Despite its simplicity, one limitation of WER is that it doesn't provide any insights into the specific nature of translation errors. Consequently, additional investigation would be necessary to pinpoint the source of the translation errors.

2.3 Annotations

In the second stage of data processing, we conducted categorical annotations on the review text, focusing on labeling sentiments and the subject of the review. Our goal was to discern specific elements or topics within the reviews to enhance the classification of customer satisfaction. The classification comprised two dimensions: sentiment and object/topic. Sentiment analysis utilized three labels — **"Positive"**, **"Negative"**, and **"Neutral"**. The latter allowed for the identification of reviews expressing both sentiments toward a single object or exhibiting a subtle sentiment (e.g., "The staff at

[3] *Motionstilbud* interactive map (Motionslisten). Last accessed: November 13, 2023

[4] Only three reviews contain only a rating and no associated text.

[5] *Hugging Face Model* (Helsinki-NLP/opus-mt-da-en)

the reception is very helpful, but the trainers are not friendly"; "The locker rooms are just ok"). For the object dimension, we employed five labels: **"Hygiene"**, **"Staff"**, **"Equipment"**, **"Location"** and **"Not Determined"** for cases where the subject was not explicitly stated (e.g., "Excellent gym, couldn't be happier"; "The prices are too high for the quality"). Annotations were performed by group members using Label-Studio, an open-source data labeling platform. To ensure consistency, an annotation guideline was drafted and used for the task; the final version is available in the Appendix A. To mitigate potential bias in labeling, contextual information, such as rating, enterprise, author, and other attributes, was excluded from the reviews provided to annotators. Reviews were randomly assigned to annotators, with a subset of 100 common reviews serving as a benchmark to assess inter-annotator agreement. The results are presented in the next section.

2.3.1 Inter-Annotator Agreement (IAA)

Since our annotations imply an entity-level sentiment analysis, based on the subject/object to which the review is aimed, we would like to ensure certain level of consistency in our annotated dataset, both for reliability of the data and to aid potential downstream tasks. To assess the reliability of our annotations, we calculate a variation of the inter-annotator agreement (IAA), reporting thus the level of consensus among the different annotators. As stated by (Hallgren, 2012) "The assessment of IRR provides a way of quantifying the degree of agreement between two or more coders who make independent ratings about the features of a set of subjects".

Since our annotation stage required five annotators, we used **Fleiss' kappa** as one of our statistical measures of the agreement as suggested by (Hallgren, 2012). Fleiss' kappa is a generalization of Scott's pi statistic, and related to Cohen's kappa statistic, with the convenience that Fleiss' kappa works for any number of raters giving categorical ratings, to a fixed number of items, at the condition that for each item raters are randomly sampled. It measures the degree of agreement in classification over that which would be expected by chance. Fleiss' kappa is defined as:

$$k = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

where:

Category	Fleiss' Kappa
Not Determined	0.46
Staff	0.80
Equipment	0.71
Hygiene	0.81
Location	0.22

Table 1: Fleiss' Kappa values for each category.

- $\bar{P} - \bar{P}_e$ gives the degree of agreement actually achieved above chance.

- $1 - \bar{P}_e$ gives the degree of agreement that is attainable above chance.

If the raters are in complete agreement then the kappa equals 1 ($k = 1$), and if there is no agreement among the raters (other than what would be expected by chance) then $k \leq 0$. The achieved IRR for a sample size of $n = 100$ was 0.04 in the first and only round. This indicates almost no agreement. Considering both dimensions, all possible combinations of unique labels makes it harder to achieve agreement among annotators. Therefore, a second round with a common discussion among annotators would have been necessary to refine the annotation guidelines. This was not done due to time constraints. Further, to gain deeper insights into the agreement levels across different categories, we also computed Fleiss' kappa for each specific category. These calculations revealed varying levels of agreement across the categories, as shown in Table 1. These category-specific kappa values highlight the differences in the degree of agreement among annotators for different categories of labels. We observe the **"Hygiene"**, **"Staff"** and **"Equipment"** categories achieving higher agreement, while **"Not Determined"** and **"Location"** having the lower agreement levels. Upon inspection, it is clear that there is little doubt when a review has a claim regarding hygiene and staff, using clear and distinctive vocabulary, such as "clean", "staff", "instructors", etc. In contrast, the guidelines for applying the "Location" label lacked clarity, reflected by the low agreement in this topic. Finally, the **"Not Determined"** label might also have been ill-defined, encapsulating too many topics that were hard to discern from each other.

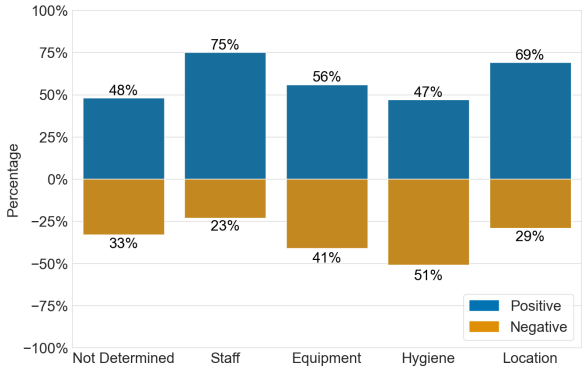


Figure 1: Distribution of sentiment across reviews (neutral excluded).

3 Experiments

3.1 Review Trends

Basic exploration of the combined annotated dataset was conducted to explore which, if any, trends existed. The sentiments for the review categories described in section 2.3, were normalised in order to explore their distribution (Figure 1); it was decided to omit visualizing the **"Neutral"** results, since they lack any polarization. For the most part, the results indicated a relatively even split between positive and negative sentiments. The most noteworthy differences from the sample of reviews was within the categories of **"Staff"** and **"Hygiene"**; with reviewers relatively more likely to positively review the staff and negatively review hygiene.

3.2 Auto labelling of incoming reviews

Our dataset holds significant potential as valuable training data for a more sophisticated review classifier, enabling the nuanced analysis of incoming reviews based on their sentiment within each category. To explore this potential, we conducted an experiment leveraging a multinomial Naive Bayes classifier for each category individually. The culmination of these individual category classifiers allows for the generation of a comprehensive multi-label classification for each review, where the final results can be appended into one multi-label classification of the review, closely mirroring the human annotation. Several transformation steps were necessary to convert the text from the reviews into a feature space for the model. To do this, we used a TFIDF (term frequency–inverse document frequency) vectorizer, that measures the importance of a word in a larger corpus of text. This approach, akin to a bag-of-words model, leverages the correlation between specific words in the review and their associated label and sentiment.

To refine this process, we implemented common pre-processing steps, including:

- **Casing and Accents:** All words in the reviews were converted to lowercase to standardize word matching, and accents were removed from words.
- **Grammar correction:** We use the library Symspellpy[6] to apply a spelling correction, on the English text, based on the Symmetric Delete spelling correction algorithm, where misspelled words were replaced with the closest matched dictionary word. The words with no match were replaced by themselves.
- **Tokenization:** Employing a word-level tokenization, each word was treated as an individual feature.
- **Stopword Removal:** Elimination of frequently occurring English words, such as articles and pronouns (refer to the Appendix B for the list of removed words).
- **Lemmatization:** We applied TextBlob[7] lemmatization with POS tagging, ensuring a more nuanced understanding of word forms. For example, verbs in different tenses or nouns in various forms were reduced to their base form (e.g., "running" to "run").

This approach not only enhances the interpretability of sentiment within specific categories but also contributes to the broader understanding of the overall sentiment landscape encapsulated in our dataset. For a more detailed discussion on the limitations of this approach, we refer to Section 4.

Since the annotated dataset is rather small ($n = 608$), we leverage the use of cross validation, applying a five-fold cross validation scheme. The average resulting metrics can be seen in Table 2. For each category, we assume one of four possible labels: **"Positive"**, **"Negative"**, **"Neutral"** or **"None"**. By including **"None"** as potential label, we emulate the human setup, where some reviews may not include a sentiment towards a specific topic.

[6] *Symspellpy Documentation (Symspellpy Docs)*
[7] *TextBlob Documentation (TextBlob Docs)*

Category	A	P	R	F1
Staff	0.66	0.37	0.33	0.31
Equipment	0.59	**0.45**	0.28	0.23
Hygiene	0.75	0.20	0.27	0.23
Location	**0.92**	0.29	0.32	0.30
Not Determined	0.63	0.35	**0.39**	**0.35**

Table 2: Performance Metrics (A: Accuracy, P: Precision, R: Recall, F1: F1-score) for Multinomial Naive Bayes Classifier Across Different Review Categories.

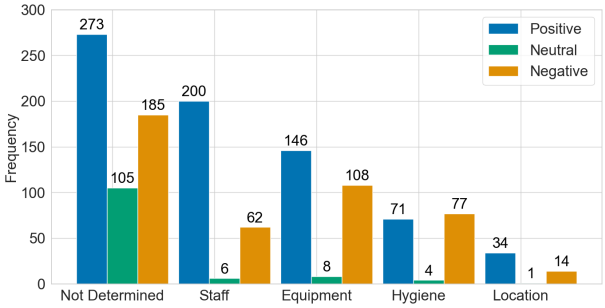


Figure 2: Absolute frequency for Positive, Negative and Neutral sentiment labels, per topic.

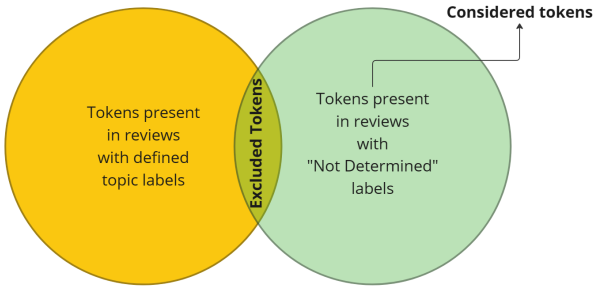


Figure 3: Diagram of tokens analysed in the "Not Determined" label.

3.3 Exploring "Not Determined" Labels

An inherent limitation in our approach lies in the predefined nature of topic labels. To ensure the relevance of these labels and assess if they encapsulate the most pertinent aspects of the reviews, we scrutinize their distribution, as illustrated in Figure 2. A closer examination, reveals that the frequency of the **"Location"** label is notably lower than the others, suggesting that this label might not be the most fitting for the dataset. In contrast, a substantial number of reviews fall under the **"Not Determined"** topic label, indicating the potential existence of crucial topics not initially targeted. To uncover these latent themes, the reviews were tokenized, and all stopwords removed. The dataset was then split into reviews that did and did not receive any kind of **"Not Determined"** label. The tokens most frequently present in reviews without a **"Not Determined"** label were identified, the implication being that they are closely associated to at least one of the defined category labels. The same was then done for the other subset of reviews, ignoring any tokens that were present in the previous subset, see Figure 3. The analysis of frequent tokens exclusively present in **"Not Determined"** labels reveals associated words[8], such as 'membership', 'pay', 'price', and 'money'. This hints that many reviews focus on the membership price of the facilities. Consequently, considering the frequency and relevance of these terms, the inclusion of a label specifically addressing membership pricing could enhance the comprehensiveness of our topic classification in future iterations.

[8] A comprehensive list of these words can be found in Appendix B.

3.4 Visualizing with Folium

For developing an interactive map, Python's Folium[9] library was used, which effectively visualizes a range of facilities across Denmark. The map employs a color-coded scheme, where markers are assigned colors based on the average rating of each facility where green indicates a high rating (≥ 4), orange a moderate rating ($\geq 2 < 4$), and red signifies a lower rating (< 2). This color scheme facilitates quick and intuitive assessment of facility quality. Additionally, the map incorporates custom icons to represent different types of activities available at each facility, such as gyms, running classes, swimming, etc. This feature adds an extra layer of clarity, allowing users to easily distinguish between the diverse range of facilities. An interactive component is integrated into the map, where clicking on any marker reveals detailed information about the facility, including its specific type/activity and rating. This interactive aspect not only enhances the informational value of the map but also improves user engagement and utility. The map is rendered in HTML format, making it optimally viewable on electronic devices. This format choice ensures both accessibility and ease of navigation for users. For a comprehensive overview, screenshots of the

[9] *Folium's Documentation* (link)

map have been included in Appendix B, Figure 2. These visual representations underscore the map's functionality and design, showcasing its value as a resource for understanding the distribution and quality of facilities throughout Denmark.

3.5 Statistical differences

Comparing Ratings Across Fitness Brands Our dataset affords us the opportunity to discern significant variations in review ratings across different fitness facility brands. A visual analysis of these distributions can be seen in Figure 4, along with average rating and standard deviation for the sample in Table 3. To rigorously examine these differences, we employed a one-way Analysis of Variance (ANOVA) test, scrutinizing the distinctions in rating means within our samples. The null hypothesis posits that all groups share the same population mean. At a confidence level of 95% (α = 0.05), the ANOVA test yielded compelling results (*Statistic* $= 103.79$, *P-value* $= 1.97e^{-64}$), leading to the rejection of the null hypothesis. Subsequently, we conducted a Tukey's Honest Significant Difference (HSD) test to evaluate the significance of differences between pairs of group means. A Bonferroni correction was applied to control the familywise error rate, resulting in an adjusted α of $\frac{0.05}{6} = 0.008$. Pairwise comparisons presented in Table 4 reveal substantial differences in mean ratings, with statistical significance observed between PureGym and SATS, PureGym and Others, as well as SATS and Others. However, caution is warranted due to disparate sample sizes. Notably, the Vesterbronx group comprises only 5 reviews, while other groups boast sample sizes exceeding of around 1500 reviews. This discrepancy may impact the robustness of observed differences and should be considered in result interpretation. Despite this limitation, the rejection of the null hypothesis for the mentioned pairs suggests that, on average, PureGym receives more favorable reviews than SATS, and, broadly speaking, facilities labeled as "others" tend to be better ranked than renowned brands.

Impact of Review Platforms on Brand Ratings To explore the potential influence of review platforms on the average ratings received by brands, we conducted a statistical analysis to identify correlations between the online platform and ratings. A Mixed Linear Model was employed, treating the online platform as an exogenous variable and rating as the dependent target variable, while considering the brand as a random effect. The results reveal that the platform Trustpilot exhibits a coefficient of 0.295. In practical terms, this signifies that, compared to the reference category (Google Maps), the mean rating is expected to increase by 0.295 when the platform is Trustpilot. The associated P-value of the test is 0.002, indicating a statistically significant impact of the choice of platform on the mean rating of the brand. Crucially, it is imperative to acknowledge that the variance of the random effect associated with the brand variable is estimated at 0.768. This estimation suggests that the variability in ratings is unique to each enterprise, highlighting significant diversity among different businesses. The higher value of this estimation underscores the substantial variability in ratings across various enterprises, emphasizing the importance of considering individual characteristics and contexts when interpreting the influence of review platforms on brand ratings. Further details of this test can be found in Appendix (Table 4).

Brand	n	Avg. Rating	SD Rating
PureGym	1,660	3.72	1.51
SATS	1,462	2.95	1.78
Vesterbronx	5	5	0.00
Others	459	4.27	1.28

Table 3: Sample size (n), average and standard deviation for rating across all brands.

Comparison	Stat.	P-value
PureGym - SATS	0.765	**0.000**
PureGym - Vesterbronx	-1.282	0.279
PureGym - Others	-0.550	**0.000**
SATS - Vesterbronx	-2.047	0.022
SATS - Others	-1.315	**0.000**
Vesterbronx - Others	0.732	0.739

Table 4: Tukey's HSD Pairwise Comparisons of mean rating across groups (95.0% Confidence Interval).

4 Discussion

In this project, our primary focus has been on the in-depth analysis of consumer reviews for fitness facilities, aiming to offer valuable insights for iden-

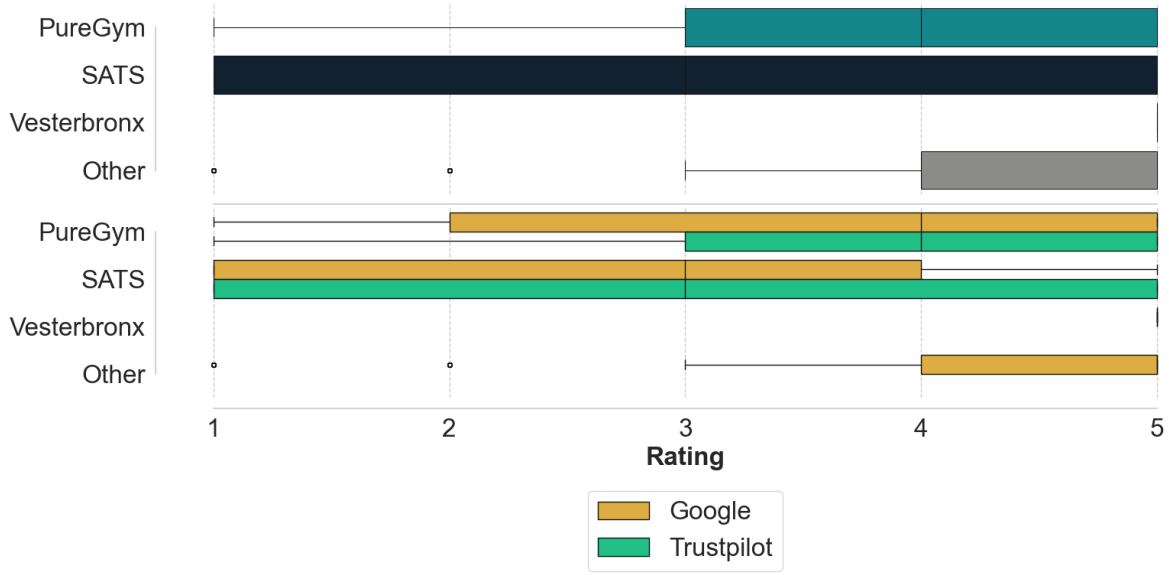


Figure 4: Top boxplot: distribution of ratings in reviews, per brand. Bottom boxplot: same distribution discriminated by review platform.

tifying potential shortcomings or pitfalls in their business operations. We present a comprehensive framework that facilitates the analysis of both historical and incoming reviews, establishing a seamless pipeline encompassing critical steps such as natural language processing for sentiment and topic analysis, brand comparison analysis, and geospatial visualization.

Our framework empowers businesses to derive actionable intelligence from the wealth of customer feedback, enabling them to strategically enhance their services and address customer concerns effectively. By integrating natural language processing techniques, we decipher the sentiment and topics within reviews, providing a nuanced understanding of customer experiences.

The brand comparison analysis within our framework allows businesses to benchmark their performance against competitors, gaining valuable competitive intelligence. Additionally, geospatial visualization enhances the spatial context of customer sentiments, aiding businesses in tailoring their strategies to specific locations.

However, it is essential to acknowledge certain limitations associated with the use of this dataset, which will be elaborated upon in the subsequent sections.

4.1 Label classifier

Our experimental framework from Section 3.2 revolves around the utilization of the multinomial Naive Bayes classifier, a well-suited model for classification tasks involving discrete features like word counts, an apt choice aligning with the nature of our application. The potential for further advancements exists through the exploration of more sophisticated models like transformers or Long Short-Term Memory (LSTM) models. These advanced architectures, known for handling complex relationships within data, could potentially elevate the performance of our classification system. Additionally, expanding our training sets presents another avenue for improvement. Enlarging the training data could contribute to enhanced model generalization and further improve overall performance, fostering a more robust and accurate sentiment analysis framework for our application.

4.2 Data scarcity

Constrained by limitations in both time and budget, our ability to curate and annotate a substantial volume of data has been restricted, consequently diminishing the depth of potential insights gleaned from the dataset. In an ideal scenario, the availability of more extensive resources would have allowed for the engagement of professional annotators. This measure would have addressed the scarcity of data, ensuring a robust and reliable training set for subsequent labeling tasks.

4.3 Bias and data quality

4.3.1 Biases in Data Itself

Halo Effect: In the realm of customer reviews, it is widely acknowledged that people are more inclined to give positive reviews to businesses that already have a good reputation, an effect especially present in the service industry, as mentioned by (Wirtz, 2003). This bias can be attributed to the prior positive experiences or perceptions consumers have about the brand.

Reviews legitimacy: There is a clear interest from businesses to obtain good reviews, as these are associated with a brand's reputation. Therefore, a shadow of doubt can be cast in the legitimacy of website reviews. Independent investigations have suggested that review websites such as Trustpilot may have fake reviews[10].

Non-response bias: An additional consideration is non-response bias, particularly pertinent in the realm of customer reviews. Recognizing that not all customers are equally motivated to leave feedback introduces an inherent limitation. Some customers may remain satisfied or dissatisfied for extended periods without expressing their sentiments unless prompted. This aspect underscores the need for a nuanced interpretation of the data, recognizing that the absence of reviews does not necessarily equate to a lack of customer experiences.

4.3.2 Collection Stage Biases

Representation bias: Our dataset's primary limitation lies in its origin from only two main sources, limiting our ability to assert that this sample of reviews accurately represents the broader population of customer sentiments toward fitness facilities. While valuable insights can be gleaned, caution is warranted in generalizing findings beyond the sources accessed.

Availability bias: The reliance on easily accessible reviews from Google Maps and Trustpilot introduces a potential bias. The prevalence of readily available data, while expedient, may not fully capture the diverse range of customer experiences. It is crucial to acknowledge that this bias stems from the inherent human tendency to favor easily accessible information.

Sampling Bias: The initial phase of our data collection process involved defining the scope of it. The absence of a clear and objective criterion for brand selection introduces a potential sampling bias, as the dataset may not fully represent the diverse landscape of fitness facilities in Denmark. The decision to include specific brands, while informed by their visibility, could inadvertently skew the findings towards the characteristics of these chosen entities. This limitation prompts a cautious interpretation of the results, recognizing that the insights drawn are contingent on the chosen sample and may not be universally applicable to the broader spectrum of fitness facilities in the country.

4.3.3 Biases in Data Processing

Misfit Bias: The pre-defined topic labels for our task (namely "Hygiene", "Staff", "Equipment" and "Location") may not fit well the actual reviews, leading to biased annotations. While "Hygiene" and "Staff" were recurrent topics in the reviews, confirming our hypothesis, "Location" was not a well defined topic, that seemed to confuse annotators and failed to accurately represent the diversity in the data.

Labeling Bias: Our examination of the labeled dataset revealed inherent biases stemming from the perspectives of annotators. As individuals, we bring our unique cultural backgrounds and personal experiences to the task of interpreting and applying labels, particularly evident in the context of sentiment analysis. The assessment of reviews was noticeably influenced by the cultural nuances and personal perspectives of the annotators, especially since some annotators are not English native speakers, introducing a layer of subjectivity. Ambiguity emerged as a substantial source of disagreement among annotators, with cultural backgrounds playing a pivotal role in shaping the final label decisions. The diverse interpretations of ambiguous content underscored the impact of individual perspectives on the labeling process. Moreover, our contextual familiarity with fitness centers introduced an additional layer of bias. Some annotators, being members of fitness centers, possessed specific contextual insights that inadvertently influenced their labeling decisions. The interplay between personal experiences and the task at hand underscored the complexity of mitigating biases in the annotation

[10] Smith, Mike Deri (January 2013). "Fake reviews plague consumer websites". *The Guardian*. The Guardian Article.

Belton, Padraig (June 2015). "Navigating the potentially murky world of online reviews". *BBC News*. BBC News Article.

Naylor, David (March 2012). "Google how can you trust www.trustpilot.co.uk ?" Bronco Article.

process, necessitating a conscientious approach to ensure the reliability and objectivity of the labeled dataset.

Inaccurate translations: The reliance on machine-generated translations for the majority of our reviews introduced a notable challenge, as the quality of these translations often fell short of perfection. This inherent limitation raised concerns about the potential compromise of the purity of the downstream annotation task. Our investigation revealed instances where inaccuracies in the translation significantly influenced the annotation process, leading to misinterpretations by annotators. Consider, for example, a review written in Danish: "Den dårligste Santa jeg har set alt den styrke du må ikke så hygiejne På Nørrebro." A human translator rendered this as "The worst sats I have seen on Nørrebro." However, when translated by the model, it translated the mispelled "Santa" to "Santa," which in this context is nonsensical and could lead to confusion among annotators. Such mistranslations highlight the critical importance of context-aware translation. They can significantly impact the data analysis process, leading to incorrect interpretations in sentiment labeling and undermining the overall accuracy of the study. It became evident that the imprecise wording of certain translations had the unintended effect of guiding annotators toward incorrect labels for specific instances. This underscores the importance of acknowledging and addressing the limitations introduced by machine-generated translations, emphasizing the need for meticulous review and validation processes to mitigate potential distortions in the interpretation of user sentiments.

5 Future Work

While substantial groundwork has been laid in this project, we are actively contemplating avenues for further development. Our primary considerations revolve around **enhancing both the quantity and quality of the data**, encompassing improvements in translations and annotations. Additionally, we intend to explore more **sophisticated natural language processing models**, as discussed in Section 4.1, to refine our understanding of customer sentiments and topic detection.

A critical aspect of our future endeavors involves a thoughtful **reassessment of the available labels** to ensure they effectively capture the diverse topics of interest within the reviews. This involves not only enriching the existing label set but also refining and possibly expanding it to accommodate the multifaceted nature of customer feedback.

Moreover, we envisage significant potential in **expanding geographical coverage.** The current focus on Danish urban areas has yielded rich insights, yet extending this scope to encompass smaller cites, rural regions and potentially other countries stands to offer a broader perspective. Such an expansion would not only diversify our dataset but also illuminate regional variations in gym reviews and customer satisfaction trends.

Another key aspect we aim to explore is the **extension of the dataset with additional review sources**. Our dataset can benefit from the inclusion of varied review platforms, such as Yelp or business websites, and even reviews posted in social media, such as Reddit or Facebook. This integration would enhance the dataset's comprehensiveness, providing a more nuanced understanding of customer feedback across different social media landscapes.

Lastly, the aspect of **reviews analysis over time** presents an intriguing prospect. Delving into how customer reviews and ratings evolve over periods can shed light on the impacts of specific events or changes within the gym facilities, such as management shifts or infrastructural renovations. Understanding these temporal dynamics is crucial for comprehending the long-term trends in customer satisfaction and their implications for the gym industry. Each of these areas not only extends the scope of our current research but also opens new doors to understanding the complex dynamics of customer feedback in the gym industry.

6 Conclusion

In conclusion, the framework developed in this project, coupled with the collected data, has yielded valuable insights into customer satisfaction within fitness facilities across major cities in Denmark. Despite the constraints imposed by the sample size, significant variations in customer satisfaction have been uncovered among major fitness brands. Notably, small "open" facilities tend to receive higher ratings, potentially attributed to their free accessibility, and PureGym exhibits on average a higher customer satisfaction than Sats.

Analysis of review platforms reveals a potential association between the platform and expected rat-

ings, with Trustpilot reviews displaying a tendency for higher ratings compared to those on Google Maps.

Moreover, the dataset has facilitated the dissection of reviews into distinct topics, offering nuanced insights into customer sentiments. Predominantly, reviews emphasize aspects related to staff and equipment, with a majority expressing positive sentiment.

In summation, this project contributes a comprehensive framework for retrieving, analyzing, and presenting customer satisfaction data for fitness brands in Denmark. While acknowledging the documented limitations, this work provides a foundation for future improvements and a more in-depth understanding of customer experiences in the fitness industry.

References

Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23.

F. Karakaya and N. Ganim Barnes. 2010. Impact of online reviews of customer care experience on brand or company selection. *Journal of Consumer Marketing*, 27(5):447–457.

S Lee, J Lee, H Moon, C Park, J Seo, S Eo, S Koo, and H. Lim. 2023. A survey on evaluation metrics for machine translation. *Mathematics*, 11(4).

Z. Mo, Y. Li, and P. Fan. 2015. Effect of online reviews on consumer purchase behavior. *Journal of Service Science and Management*, 8:419–424.

Jochen Wirtz. 2003. Halo in customer satisfaction measures : The role of purpose of rating, number of attributes and customer involvement. *International Journal of Service Industry Management*, 14:96–119.

A Annotation Guide for Review Sentiment and Topic Classification

Sentiment Annotation: In this annotation project, our primary objective is to assess the sentiment of reviews, categorizing them as either **POSITIVE**, **NEGATIVE**, or **NEUTRAL**. We employ a simple, one-layer annotation system, ensuring mutually exclusive labels. Annotators are instructed to assign **"POSITIVE"** for reviews expressing a clear and definite positive view, **"NEGATIVE"** for those with a discernible negative tone, and **"NEUTRAL"** for reviews lacking a clear positive or negative stance.

This approach aims to effectively address mixed sentiment reviews. In instances where a review contains both positive and negative elements, annotators should select both sentiment labels. If both contradictory sentiments are guided towards the same topic, this combination will be addressed as **"NEUTRAL"**. It's important to highlight that the **"NEUTRAL"** label cannot be directly applied in the Label Studio setting, and thus both **"POSITIVE"** and **"NEGATIVE"** labels should be selected instead.

Topic Annotation: For topic annotation, we utilize a one-layer scheme with four possible and potentially overlapping labels: **"EQUIPMENT"**, for reviews related to gym fitness equipment; **"LOCATION"**, for those focusing on the geographical location of the gym; **"HYGIENE"** for reviews concerning the cleanliness of the facilities; and **"STAFF"** for those centered around the personnel running the facilities. The **"NOT DETERMINED"** label serves as a fallback option when none of the four available topics (Equipment, Hygiene, Staff, Location) are applicable to a review. While the **"NOT DETERMINED"** label is not explicitly selectable in Label Studio, the presence of a sentiment label with no topic label will be rendered as **"NOT DETERMINED"**. It's essential to recognize that all labels can be overlapping. This means that more than one label can be chosen for each review if multiple aspects are relevant, each with a potentially different sentiment.

Preliminaries
Below, we introduce the labels of this annotation project. It takes only a few minutes to read. A brief explanation of the Subject classes:

1. **Equipment**: This refers to specific or general aspects of the physical machinery installed in the gymnasium. More specifically, it refers to the various tools, machines, devices, and gear used by individuals to engage in physical exercises and activities. Examples of equipment include machines like treadmills, stationary bikes, free weights (dumbbells, barbells), resistance bands, medicine balls, and overall accessories, like weightlifting gloves, lifting belts, etc.

2. **Location**: This refers to the actual geographical location of the fitness facility, as well as its surroundings and ease of access. More specifically, it could include things like neighborhood, nearby public transportation, parking spaces, etc.

3. **Hygiene**: This refers to the overall maintenance of a facility's cleanliness and order. It could also be extended to things like how easy it is to find the equipment, the state of bathrooms and showers, lockers, etc.

4. **Staff**: This refers to the staff who work around the facility. More specifically, here we look at their competence and demeanor, both of which influence the overall gym experience.

5. **Not determined**: For those cases where the subject of the review is not directly and/or clearly stated.

```
1.Author: Ahmad Kasem Haidar, Rating: 3, Review: Englandsvej, It's good and I enjoy training there with different
machines BUT I didn't give 5 stars because we have to refill our water bottles from the toilet sink!
Hope they will make a separate place for water. --> end review.

2.Author: Sebastian, Rating: 1, Review: Yet another Fitness World center, which have been worsened by the PureGym-
makeover.
Strenght training has been compressed into too little space and instead they added a lot of useless cardio equipme
nt for amateurs. It's becoming more and more clear that the latter is the clientele which PureGyms wants to cater
to - not us, who consistently go to the gym +5 times a week.
It's such a pity with this systematic destruction of the relatively few good gyms we have. --> end review.

3.Author: horia cunea, Rating: 5, Review: Great place and great staff, also Victor was very helpful ans answered a
ll my questions very nice --> end review.

4.Author: Mikhail Nikitin, Rating: 4, Review: After renovation they made an "open office" environment in the chang
ing room. Bizarre. And still no soap or shampoo in the shower. --> end review.

5.Author: Jennifer C, Rating: 5, Review: Excellent gym. Many locations, good value for money, friendly staff, well
equiped and huge variety of classes. --> end review.
```

Figure 1: Example of reviews to annotate.

Annotation example:

Labels:

1. Positive, Equipment; Negative, Hygiene
2. Negative, Equipment
3. Positive, Staff
4. Negative, Hygiene
5. Positive, Location, Staff, Equipment

Thank you for your careful and precise annotations, which contribute significantly to the quality of our sentiment and topic classification.

B Tables and Figures

Complete dictionary of removed tokens (Stopwords) for classifier pre-processing: am, seemed, full, hereby, becoming, anyhow, seems, next, already, herself, me, made, four, up, much, that, off, became, then, will, either, side, more, except, whole, most, see, with, go, everyone, were, your, rather, so, some, nothing, thus, itself, de, beyond, this, although, too, why, via, amongst, put, sometimes, latterly, co, less, etc, mine, thereupon, i, interest, nobody, any, becomes, below, well, couldnt, otherwise, within, there, him, such, anything, she, forty, myself, whence, who, eg, cry, my, whether, ours, even, beforehand, top, how, thereafter, often, fill, these, every, what, serious, thence, them, together, about, once, whereafter, own, none, six, sixty, not, system, become, across, anywhere, he, several, someone, almost, sincere, never, therefore, many, and, somewhere, whoever, thick, whom, per, here, to, former, his, found, each, upon, however, ever, mostly, could, inc, mill, on, give, thru, hundred, toward, further, can, they, though, during, around, again, anyone, because, noone, bill, without, due, alone, been, call, something, it, first, a, has, would, one, others, than, detail, was, where, if, yours, everything, while, hereupon, for, their, meanwhile, find, please, nine, is, her, by, from, which, do, all, else, fifteen, are, should, sometime, very, beside, had, or, bottom, elsewhere, the, us, before, empty, among, being, fifty, now, anyway, another, ten, front, ie, few, name, since, must, may, thereby, seem, two, yourself, we, take, eleven, done, as, everywhere, whereupon, nor, but, con, between, you, through, therein, themselves, latter, down, those, above, yet, three, whereby, least, over, its, until, hereafter, yourselves, no, out, behind, be, might, besides, wherever, hence, onto, along, thin, have, namely, third, enough, whenever, back, show, against, indeed, formerly, hers, whereas, in, also, hasnt, under, eight, still, of, himself, always, cant, throughout, last, get, whither, other, twenty, part, when, afterwards, perhaps, moreover, un, nevertheless, nowhere, ltd, keep, towards, neither, re, whatever, cannot, five, somehow, whose, describe, twelve, herein, seeming, move, amoungst, into, only, wherein, at, same, our, an, fire, amount, after, both, ourselves.

List of most frequent words found exclusively in "Not Determined" labelled reviews: gym, good, staff, machines, nice, center, training, equipment, place, one, always, great, fitness, clean, also, many, people, really, get, time, like, service, friendly, well, bad, train, room, membership, could, sauna, month, need, space, everything, work, classes, best, music, better, months, years, team, free, workout, atmosphere, super, area, day, old, pay, changing, find, often, weights, price, floor, broken, 2, around, gyms, cleaning, facilities, rooms, say, several, however, hours, much, centers, customer, open, copenhagen, money, large, experience.

Field	Description
datetime	Timestamp of the review.
name	Review author.
rating	Rating given, from 1 to 5.
title	Title of the review.
review	Review text.
event_time	The date of the event/experience that the review refers to.
enterprise	Name of the company being reviewed.

Table 1: Data fields collected from Trustpilot website.

Field	Description
Type	Activity category.
Name	Name of the activity, and where it takes place (if available). E.g. "Kondisti hos Valby-parken".
Contact Information	Any combination of the following: Link to website, telephone number, email address.
Gender	Target group gender: Male, Female, Both.
Age Group	The target group age (e.g. "Seniors").
Special Target Group	Specified if there is an intended target group (e.g. "overweight" or "65+").
Meeting Place	The address of the intended activity.

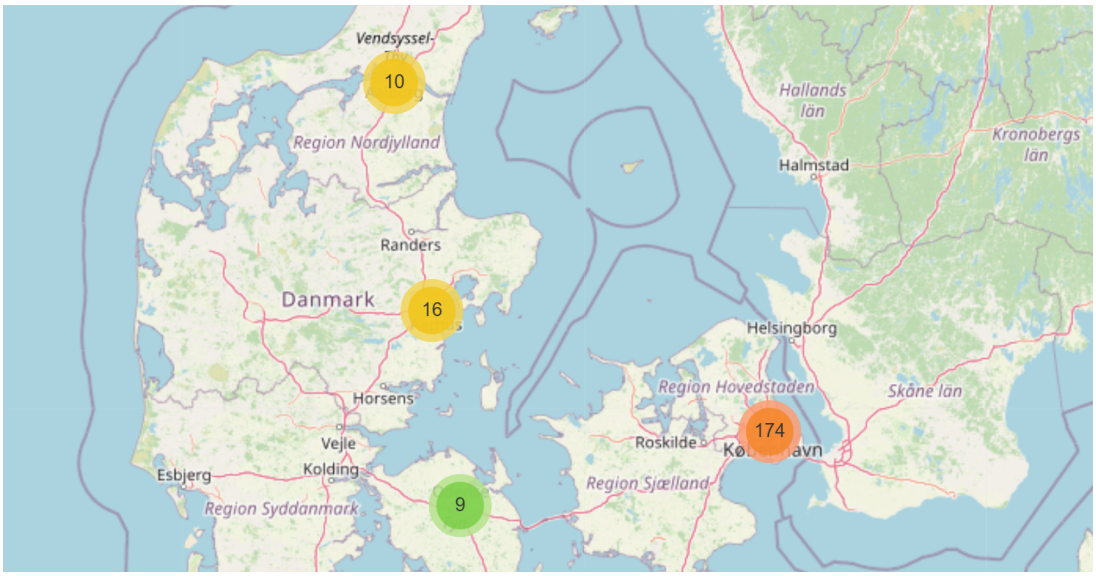
Table 2: Collected attributes from Copenhagen Municipality's Motionstilbud interactive map.

Field	Non-Null Count	Data Type	Description
enterprise	3586 non-null	String	Brand or business ("OTHER" if not in our target group)
author	3586 non-null	String	Author of the review.
rating	3586 non-null	Integer	Stars or rating given in the review (scale from 1 to 5)
review	3583 non-null	String	Review text, if present.
platform	3586 non-null	String	Platform where the review was posted (either Google or Trustpilot)

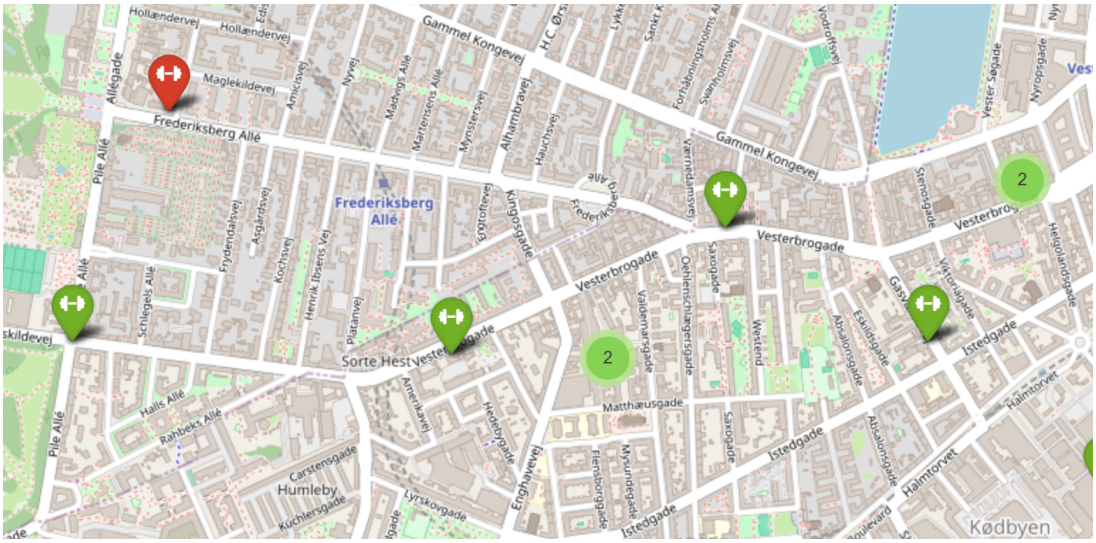
Table 3: Detail of attributes for final dataset.

Dependent Variable	Rating
# Observations	3,586
Groups	PureGym, Sats, Vesterbronx, Others
Min. Group Size	5 (Vesterbronx)
Max. Group Size	1,660 (PureGym)
Avg. Group size	896.5
Intercept	3.725 (mean rating)
Trustpilot	0.295 (effect) (**P-Value: 0.002**)
Enterprise	0.768 (effect)

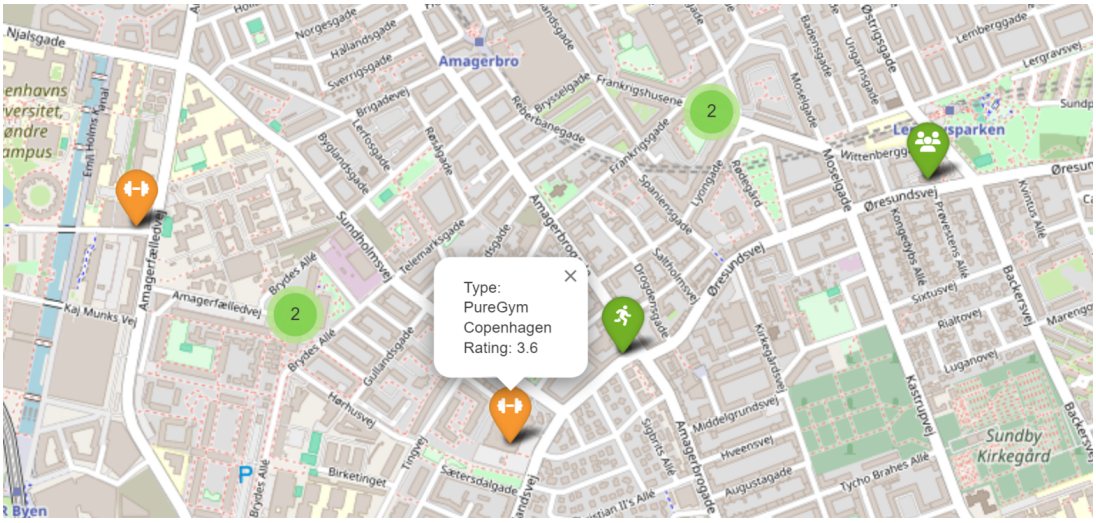
Table 4: Mixed Linear Model. Ratings as dependent variable, platform as exogenous and enterprise as random effect.



(a) Overall map view.



(b) Close-up look of the map.



(c) Close-up look with toolbox information of the place.

Figure 2: Screenshots of the interactive map of the reviews.