Copenhagen Business School

Master of Science in Business Administration & Data Science
Natural Language Processing and Text Analytics

Supervisors: Daniel Hardt & Rajani Singh

Final Project

Unleashing the Power of Large Language Models: GPT-3.5 and BERT versus Traditional Models for Sentiment Analysis in Airbnb Reviews

Dataset and Code

https://drive.google.com/drive/folders/1NRPdL-0tuVwb-APwUQS7H6r8XFuICiCc?usp=share_link

Authors	Konstantin Kleffke, Linda Schombach, Julius Wirbel & Yile Huang
Student IDs	158284, 158305, 158289, 158299
Pages	15
Characters	33,790
Submission	30th of May 2023

Abstract

Large Language Models (LLMs) gain increasing attention and offer new potential in the field of natural language processing. Companies such as Airbnb could leverage those models to extract valuable information. This paper aims to develop and compare two of those LLM models (BERT and GPT-3.5) with traditional models (Logistic Regression and Naïve Bayes) for binary sentiment analysis based on scraped Airbnb review comments. It was hypothesized that LLMs achieve better performance due to their advanced linguistic understanding and contextual comprehension. Results show that GPT-3.5 has the highest macro F1 score of about 0.906. Surprisingly, both Logistic Regression and Naïve Bayes perform similarly or even better than the BERT model. It is concluded that GPT-3.5 could offer great benefits for Airbnb. However, ethical implications and limitations must be considered. Further, this paper identifies important customer topics using LDA and Top2Vec which include location, amenities, host, recommendation, and booking.

Keywords: *Natural Language Processing, Large Language Models, Sentiment Analysis, GPT-3.5, BERT, Topic Modeling, Latent Dirichlet Allocation, Top2Vec, Airbnb*

Table of Content

Table of Figures

Table of Tables

1. Introduction

Airbnb is an online platform offering 6.6 million short-term rental accommodations worldwide (Airbnb, Inc., 2022). Acting as an intermediary, it connects people who want to rent their private housing and guests. Airbnb's vision is that "hosts offer unique stays and experiences that make it possible for guests to connect with communities in a more authentic way" (Airbnb, Inc., 2022).

Considering this vision as well as the risk associated with private house rentals, ensuring a functioning and reliable quality check is of utmost importance. In the digital area, electronic Word-of-Mouth (eWOM), such as online hotel reviews, has become an increasingly important source of information as access is easier and unrelated to personal connection (Liu, 2006). Due to the inherent risk and uncertainty of travel-related decision-making this becomes relevant (Chiny et al., 2021; Mellinas & Reino, 2019). Also, eWOM information is perceived more credible than traditional advertising (Weyerer, 2019).

However, the rapid growth of reviews makes it challenging to analyze this vast amount of data. To ensure that the insights present in the review data are leveraged in the best possible way, natural language processing (NLP) techniques can be employed (Raza et al., 2022; Rezazadeh et al., 2021; Von Hoffen et al., 2018). Recently, Bidirectional Encoder Representations (BERT) and Generative Pretrained Transformer (GPT) marked a breakthrough in the accuracy and training approach taken for deep learning models in NLP (Ray, 2023). The former was developed by Devlin et al. (2019) at Google. The latter was launched by OpenAI in 2020 (fine-tuned to GPT-3.5 in 2022). It is based on reinforcement learning from human feedback and aligns the model to human preferences (Wang et al., 2023).

This paper aims to address the challenge of increasing text of Airbnb reviews by developing and comparing different models. Results could support Airbnb in choosing the right model and in providing better information for customers. The evaluation of a review being positive or negative seems to be most important for the customer and their purchasing decision. Thus, sentiment analysis was chosen which is a classification task in NLP that seeks to identify underlying opinions, sentiments and emotions expressed towards an entity (Medhat et al., 2014; Wang et al., 2023).

It is hypothesized that large language models (LLMs) such as GPT and BERT will outperform more traditional learning models like Logistic Regression and Naïve Bayes (NB). This is based

on the expectation that the LLMs exhibit a higher level of understanding of the reviews' linguistic complexities. They are aware of the context, and as a result, can capture subtle nuances and detect sentiments based on context, tone, and implied meanings. This ability is particularly beneficial for sentiment analysis tasks, where the sentiment can often hinge on the understanding of context and subtext.

The paper begins by outlining the process of generating and pre-processing a dataset comprising over one million Airbnb reviews. Subsequently, an exploration of Airbnb customer needs is conducted through the development of topic models, namely Latent Dirichlet Allocation (LDA) and Top2Vec. As baseline models, NB and Logistic Regression are employed and for the LLMs, BERT model in conjunction with a custom classifier and GPT-3.5.

2. Related Work

Various research exists in the field of Airbnb reviews. Particularly due to accessibility of reviews on their website due to scraper tools opens diverse research areas ranging from sentiment analysis to topic modeling (Guttentag, 2019). To begin with, online reviews can be explored to identify topics customers care about in the Airbnb domain. Cheng and Jin (2019) found four major topics in online reviews: location, amenities, host, and recommendation. Luo and Tang (2019) uncovered five aspects (communication, experience, location, product/service and value) as well as two emotions (joy and surprise) as most important in Airbnb reviews using Latent Aspect Rating Analysis (LARA).

Considering sentiment analysis, Lawani et al. (2019) conducted a study on factors affecting the pricing of Airbnb accommodations and found a positive relationship of positivity of reviews and the price. They used a lexicon-based model to calculate the sentiment of a review by summing the positivity or negativity of each word. Similarly, Rezazadeh et al. (2021) utilized sentiment analysis with TextBlob to extract features from Airbnb reviews which was then used to predict pricing. Other methods of analyzing reviews range from the traditional NB classifier (Khomsah, 2020) to deep learning approaches (Raza et al., 2022). Raza et al. (2022) compared Recurrent Neural Network, Long Short-Term Memory, and Gated Recurrent Unit methods for sentiment classification of Airbnb reviews. Results support that Gated Recurrent Unit has the highest accuracy rate and other performance indicators.

Considering advanced deep learning models, ChatGPT (build upon GPT-3.5) was compared by Zhong et al. (2023) with four representative fine-tuned BERT-style models. While their

findings suggest that ChatGPT is not performing comparably well in handling paraphrases and similarity tasks, it outperforms the trained BERT models on inference tasks. Regarding sentiment analysis, both models achieve a similar performance. Similarly, Wang et al. (2023) found that ChatGPT achieves similar results as fine-tuned BERT for sentiment classification. In conclusion, those studies suggest that Airbnb review sentiment analysis could perform similarly or even better using GPT-3.5 compared to BERT.

3. Methodology

3.1. Dataset Description

To ensure the usage of a dataset that is up-to-date and best suitable for the present research question, datasets were scraped via Apify. Apify is a website providing a variety of ready-made actors for web scraping. By defining location, price range, and other needed variables, several datasets were scraped sequentially using the Airbnb Scraper. Due to the limited capacity of returning datapoints, it was required to request data in several iterations. It was scraped based on a selection of cities and by different price ranges. Location was limited to Europe. The seven cities with the largest average monthly profit were selected as they were considered to have sufficient datapoints: London, Paris, Munich, Dublin, Milan, Copenhagen and Oslo (Rokou, 2022). Overall, the dataset prior pre-processing holds 16,106 Airbnb's and 1,077,494 reviews.

3.2. Data Pre-Processing

The dataset is pre-processed as visualized in Figure 1. After loading the data, it was merged into one dataframe containing information about all scraped Airbnb's. After several pre-processing steps for all models which will be explained in the following, some additional steps were applied. Top2Vec, BERT and GPT-3.5 were trained without those additional steps.

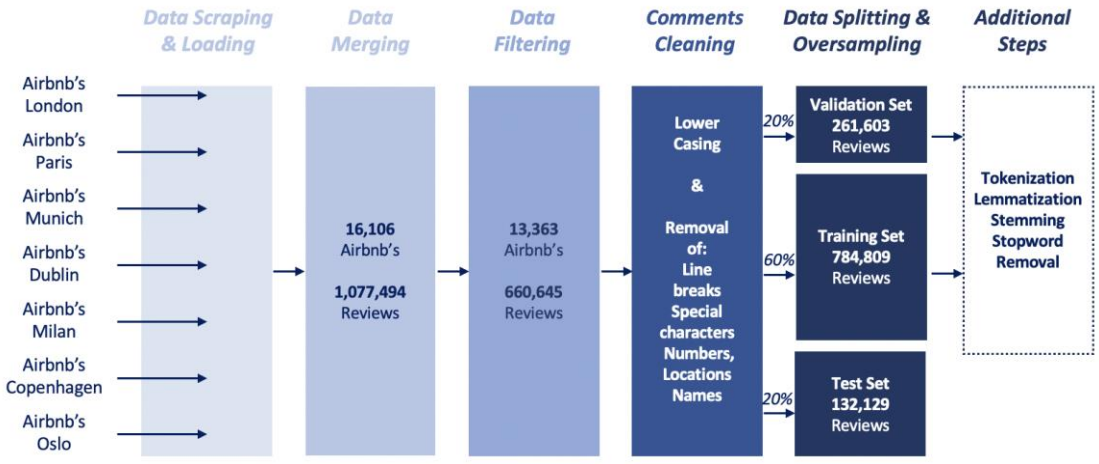


Figure 1: Pre-processing Pipeline

3.2.1. Data Filtering

First, duplicates were checked, which is especially important as the data was scraped using several extraction loops. This was done based on the same Airbnb's URL and review IDs. In addition, Airbnb's with the same name and location were removed. To have a consistent language, the language of the reviews was checked. Non-English comments were not translated but removed as they provide a sufficient number, and a long running time could be avoided. Airbnb's and reviews having null values were removed. Besides null values, outliers can falsify the results. Several visualizations were created and inspected to detect outliers. Especially the price per guest revealed some unrealistic entries that were deleted. For instance, the Airbnb's in Figure 2 were concluded to be fake postings or wrongly scraped. In conclusion, all entries with a price per guest above 400 USD were removed.

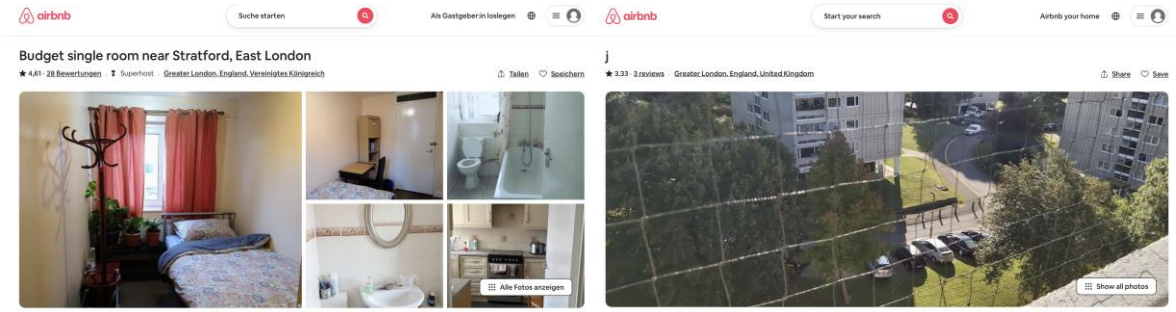


Figure 2: Examples of identified outliers by price per guest (link Airbnb 1, link Airbnb 2)

3.2.2. Review Comments Cleaning

Line breaks, special characters, numbers, locations, and names were removed from the text. For instance, the location does not contribute substantially to the sentiment analysis on a broader level. Also, all words were converted into lower cases.

3.2.3. Data Splitting & Oversampling

As displayed in Figure 3, there was a high imbalance of review ratings. 559,724 comments, which corresponds to over 81% of all reviews, were rated with the top score of 5.0. It was decided to use binary ratings (positive and negative). In addition, all reviews containing a rating of 3.0 were removed as they were not attributable to one of the binary classes but had a neutral tone. Further, zero rating reviews were deleted as they are synthetically created by Airbnb and thus not reliable (Lee et al., 2020).

This results in 654,007 positive (5.0 and 4.0) and 6,638 negative ratings (2.0 and 1.0) as shown in Figure 3. The imbalance created by the overwhelming amount of positive feedback poses a challenge for good performing sentiment analysis. Bridges & Vásquez (2018) as well as Alsudais and Teubner (2019) identified a notable bias towards positive reviews in online

feedback. The latter noted that only 1.06% of reviews they analyzed were negative. Also, Santos et al. (2020) found that particularly the economy Airbnb is in causes more positive reviews. Due to shared or at least private accommodations, Airbnb hosting's are likely to create personal relationships and thus more positive feedback.

The dataset was randomly divided into a training (60%), validation (20%), and test (20%) set. To balance the distribution of the binary classes, oversampling was used for the training and validation set to increase the number of negative reviews synthetically. It was tested between two oversampling techniques, random oversampling and SMOTE. Random oversampling duplicates randomly selected vectors from the underrepresented class (Glazkova, 2020). Despite its simplicity, it generates similar or better results, also depending on the model, when handling text data compared to more advanced methods such as SMOTE (Glazkova, 2020). For each model, the best performing oversampling technique was chosen.
The final training set contains 784,809 and validation set 261,603 reviews with approximately balanced classes. The test set remains without oversampling resulting in 132,129 reviews.

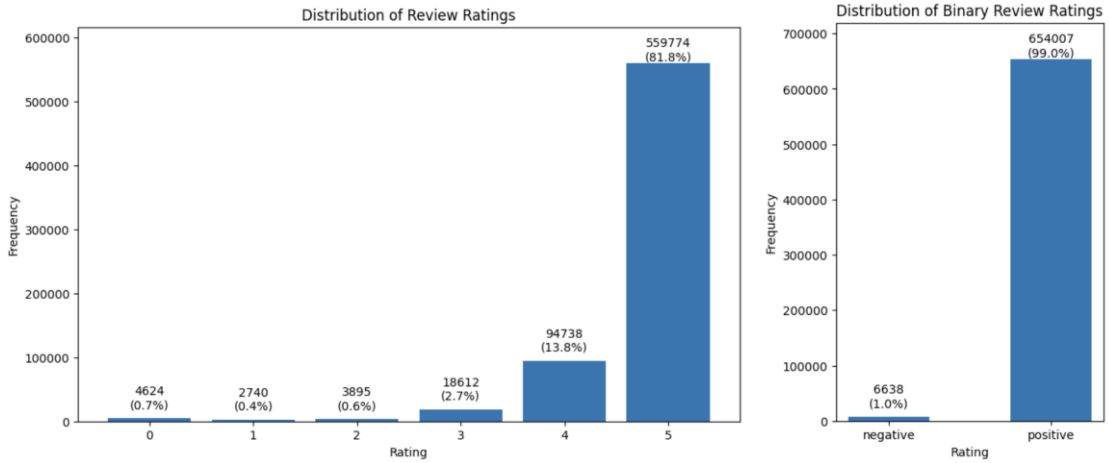


Figure 3: Distribution of review ratings without oversampling

3.2.4. Additional Steps

For some models further processing is necessary. To enhance the performance of the machine learning algorithms, lemmatization and stemming was applied to decrease the number of word variations and normalize the text data (Jurafsky and Martin, 2009). By aggregating similar words and reducing them to their base or root forms, the core semantic content is kept while minimizing redundancy. Furthermore, stopwords were removed to eliminate commonly occurring words that provide little discriminatory information and focus more on the meaningful terms. However, some negators such as 'not' were not removed. A filtering step was implemented to remove infrequently occurring words that appeared less than 10 times. Lastly, the reviews were tokenized into unigrams, bigrams, and trigrams.

Consequently, the dimension of the document-term matrix is reduced aiming to strike a balance between keeping meaningful words and removing words that do not contribute substantially for sentiment analysis. Also, the resulting vectors for each review are less sparse. The performance of the classifiers is enhanced, and the computational efficiency is improved.

3.3. Topic Modeling

To get a better understanding of relevant topics included in the review comments, LDA was employed. Following the Dirichlet distribution, LDA assumes that all reviews contain a mixture of topics, and each topic is a mixture of words (Blei et al., 2003). One limitation of LDA is the requirement to know the number of topics in advance. To address this limitation, the analysis was conducted with a varying number of topics. Also, LDA was conducted using two variations of representations: a bag-of-words (BoW) representation with and without tf-idf weighting. The tf-idf weighting scheme assigns importance to each token in a review based on its frequency within the review and its prevalence across the entire corpus of reviews capturing the relative importance of tokens (Blei et al., 2003). The best result was obtained with six topics. However, only five were considered meaningful (Table 1). The topic 'recommendation' appeared less often. It is important to note that recommendations are often derived from positive experiences, influenced by factors such as the host's hospitality, and should not be evaluated in isolation.

Topics (tokens appearing in %)	Sample words
Host (27.6%)	great, stay, host, nice, clean, help, everything, communication
Amenities (20.9%)	room, bed, kitchen, small, bathroom, bedroom, space
Booking (19.6%)	time, go, visit, want, trip, book
Location (17.3%)	walk, close, restaurant, station, area, neighborhood
Recommendation (3.6%)	enjoy, anyone, definitely recommend, pleasant, thanks much

Table 1: Topics identified with LDA

For further analysis, the Top2Vec model was applied addressing some of the LDAs drawbacks. For example, it is not necessary to know the number of topics in advance. Unlike LDA, Top2Vec does not learn from a word distribution but utilizes joint documents and semantic word embeddings for learning (Angelov, 2020). Identifying similarities in the semantic space, Top2Vec identifies topics and their corresponding words effectively. Compared to LDA, Top2Vec does not assign a mixture of topics to one review but only one. Further minor topics discovered by Top2Vec can be seen in Table 2.

Topics (assigned to reviews in %)	Sample words
Cleanliness (0.4%)	dirty, stain, dust, mold, cleaned, sheet, smell
Accessibility (0.3%)	stair, narrow, steep, staircase, mobility, climb, carrying, heavy
Air conditioning (0.2%)	conditioning, ac, fan, hot, heatwave, air, temperature, cooling
Appliances (0.2%)	dishwasher, dryer, washer, machine, microwave, toaster, iron

Table 2: Additional topics discovered with Top2Vec

The predominant topic in Top2Vec was ‘recommendation’ (0.9%) as evident from the inclusion of the word ‘thanks’ and thus deviating from the results of LDA (Figure 4).

Figure 4: Wordcloud of ‘recommendation’ topic from Top2Vec

3.4. Modeling Framework

This study focuses on predicting the negative or positive sentiment in Airbnb reviews based on training data with pre-existing sentiment labels. NB and Logistic Regression are chosen as baselines due to their simplicity and distinct modeling techniques which will be explained in more detail later. Utilizing the BoW approach, testing was conducted to evaluate the inclusion of unigrams, bigrams, and trigrams. The baseline is compared to BERT and GPT.

3.4.1. Naïve Bayes

NB uses the Bayes’ theorem to classify text. During the training phase, NB calculates the probability of each unique word for each class based on the observed frequencies in the training data to make a prediction. A grid-search with 5-fold cross-validation is conducted to select the optimal smoothing parameter. It was found that Laplace smoothing resulted in the best performance, which adds one to the count of each word in the training corpus to handle unseen words (Jurafsky and Martin, 2009).

3.4.2. Logistic Regression

Logistic Regression assigns a weight to each feature and adds a bias term. The resulting score is mapped utilizing the sigmoid function to assign a class to each review (Jurafsky and Martin, 2009). It was tested using grid-search with 5-fold cross validation to select the optimal regularization parameter to account for overfitting.

3.4.3. Bidirectional Encoder Representations from Transformers (BERT)

For the deep learning model, the DistilBERT adaptation by Sanh et al. (2020) was chosen. It was selected due to its good performance on many general NLP tasks and its ease of adapting it to the specific task utilizing transfer learning techniques. A custom classifier was trained on top of the features extracted from the base DistilBERT model.

Similarly, to GPT (Radford et al., 2018) and ELMo (Peters et al., 2018), the BERT model was pretrained on two extremely large language datasets. The authors used the BooksCorpus Zhu et al. (2015) and the English Wikipedia pages, resulting in over 3,000 million words. Unlike GPT and ELMo, BERT was trained in a way that served whole sequences instead of single words to better learn the language structure (Devlin et al., 2019). Moreover, the pretraining of the BERT model differs in two further important ways:

Firstly, it uses a masked language model (MLM) that was inspired by language modeling research from the 1950s (e.g., Taylor, 1953). Instead of showing the model all tokens of the input at any given time, 15% of tokens are randomly blocked at each iteration aiming to predict these masked tokens based on the context (Devlin et al., 2019).

Secondly, to learn the relationship between sentences, the model is trained for Next Sentence Prediction (NSP). This is achieved by taking example sentences from the input dataset and trying to predict if sentence B follows sentence A (Devlin et al., 2019).

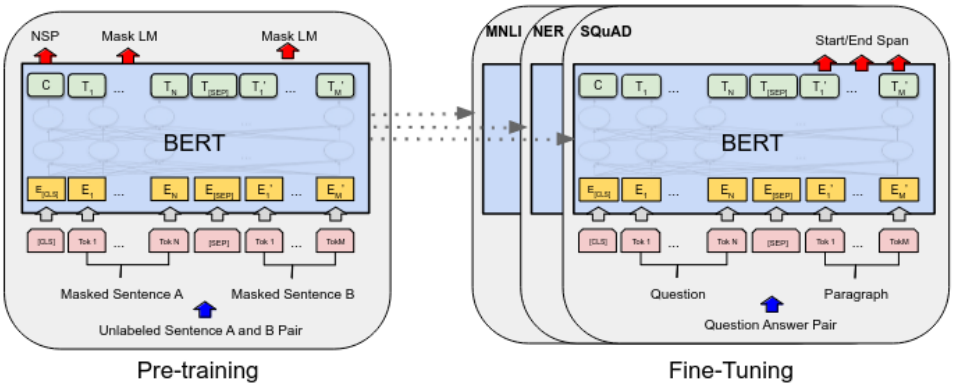


Figure 5: BERT model architecture (Sanh et al. (2019), p.3)

By distilling the BERT model following the approach laid out by Hinton et al. (2015), Sanh et al. (2020) managed to reduce the size and complexity creating DistilBERT. The network size was reduced by training a smaller 'student' network to mirror the behavior of a larger 'teacher' network. In the present case, the teacher is the original BERT-base model with 110 million parameters (Devlin et al., 2019), while the student model has half the number of layers and does not have the token-type and pooling layers present in BERT (Sanh et al., 2020). This results in a model with 66 million parameters that retains 97% of the original model's performance (Sanh et al., 2020).

3.4.3.1. Model Creation

The configuration of the classifier follows a standard classifying methodology of reducing the number of neurons towards the end of the network. To optimize the performance on the GPU, the number of neurons is always a multiple of two. The number of layers and number of

neurons per layer was determined during a manual design process. While a classifier with the same number of neurons per block was tested, it performed slightly worse than reducing the number of neurons by 50%. The dropout rate, activation function, and kernel initialization were determined by a grid search. The tanh activation, while giving up small amount of precision, performs slightly better on recall and F1 score and was thus chosen over the more commonly used ReLU activation.

Layer Name	Type	Activation	# Parameters	Initialization
input_1	Input	None	0	-
dense	Dense, 256	Tanh	196864	he_normal
dense_1	Dense, 128	Tanh	32892	he_normal
dropout	Dropout, 10%	None	0	-
dense_2	Dense, 64	Tanh	8256	he_normal
dense_3	Dense, 32	Tanh	2080	he_normal
dropout_1	Dropout, 10%	None	0	-
output	Dense, 2	Softmax	64	he_normal

Table 3: BERT model architecture

3.4.3.2. Finetuning

To adapt the DistilBERT model for classifying the sentiment of Airbnb reviews, a custom classifier based on the features extracted from DistilBERT was created. The initial approach was inspired by Feldges (2022). However, to save training time and develop the classifier faster, the features were extracted using the DistilBERT model beforehand. As the weights within the DistilBERT model are frozen and therefore not updated during training, this approach is viable and yields the same result as if the feature extraction was performed during training. To ensure the best model from the training, each time the validation loss improves, the current model weights are saved. The training configuration is shown in Table 4.

Parameter	Setting
Optimizer	Adam optimizer
EarlyStopping	64 epochs
Learning Rate Reduction	lr*0.5 after 32 epochs

Table 4: BERT model configuration

The steps visible in the Figure 6 is due to the reduction of the learning rate by multiplying it with 0.1, resulting in a smaller step size. These smaller steps lead to a better gradient update towards the minimum.

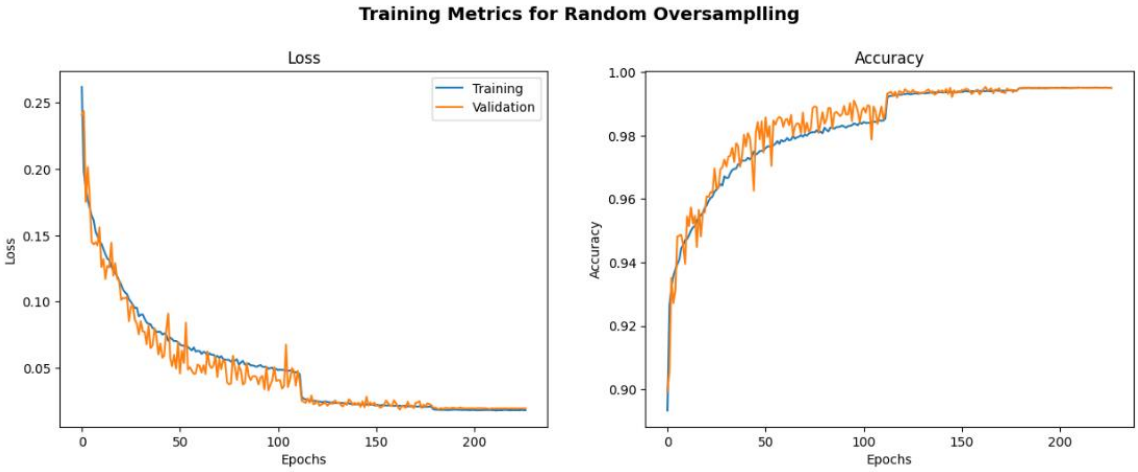


Figure 6: BERT model training metrics

3.4.4. GPT-3.5 "text-davinci-003"

GPT-4 is the most recent and advanced GPT model, but "for many basic tasks, the difference between GPT-4 and GPT-3.5 is not significant" (*OpenAI API*, 2023). Thus, this paper leverages the GPT-3.5 "text-davinci-003" model, which is a particularly specialized variant of GPT, equipped with 175 billion parameters. In this study, the performance was evaluated using a randomly selected sample of 1,000 reviews from the test set. The choice of this sample size was primarily dictated by the cost associated with making API requests to OpenAI.

This model, like the BERT model, is pre-trained, which negates the need for explicit training with a sentiment-labeled review dataset. Instead, it capitalizes on its pre-training to interpret the sentiment of text based on patterns learned from an extensive dataset. The model is prompted with the following to analyze each review and predict the sentiment:

Given the following Airbnb review, predict the reviewer's rating as 1 or 0, where 1 is positive and 0 is the negative. Provide your answer as only an integer. Here is the review: '{comment}'

The predicted sentiment was parsed from the model's response and recorded. The process was implemented with *retry* to manage possible errors or timeouts during the API calls. Once all reviews were processed, a dataframe was generated containing the review comment, the predicted sentiment, and the actual sentiment for each review. This dataframe served as the basis for evaluating the performance.

4. Results

4.1. Comparison of Models

Results of the different models are presented in Table 5. All models have a similar accuracy of about 0.99. However, given the significant imbalance in the dataset, it's crucial to look beyond accuracy when evaluating and comparing the performance of different models. Simply predicting positive sentiment each time could result in high accuracy due to the skewed distribution of the data, but this approach would lack meaningful insight and predictive power (Kulkarni et al., 2021). Therefore, precision and recall are important metrics as they also inform about the ability to correctly identify each class and avoid false positives. The F1 score, a harmonized means of precision and recall, makes performance differences apparent.

Naïve Bayes (Macro $F_1 = 0.815$)			
Label	Precision	Recall	$F_1 Score$
Negative	0.527	0.798	0.635
Positive	0.998	0.993	0.995
Logistic Regression (Macro $F_1 = 0.836$)			
Negative	0.594	0.785	0.676
Positive	0.998	0.995	0.996
BERT Model (Macro $F_1 = 0.788$)			
Negative	0.788	0.464	0.580
Positive	0.991	0.998	0.994
GPT Model (Macro $F_1 = 0.906$)			
Negative	0.750	0.900	0.818
Positive	0.997	0.991	0.994

Table 5: Model results

Overall, GPT outperformed all the other models with a macro F1 score of about 0.906. The GPT model demonstrates a strong balance between precision (0.750) and recall (0.900) and addresses the weaknesses of the other trained models in this paper.

Further, Logistic Regression slightly outperforms NB with a macro F1 score of 0.836 compared to 0.815. The more complex BERT model obtains the lowest macro F1 score of 0.788. There is no significant performance difference observed for the positive while there is for the negative class: The simpler models, NB and Logistic Regression, tend to have a low precision with 0.527 and 0.594, respectively. They compensate for this with a high recall value of 0.798 and 0.785 respectively, capturing a large proportion of the true negative class. The BERT model, however, performs better in terms of precision (0.788) while its recall is lower (0.464).

Even though the BERT model has a lower macro F1 score compared to the simpler models, BERT should not be considered inferior. The higher precision value indicates that the model

understands the underlying features of the negative class better and consequently has more confidence when predicting this class. This is one aspect that the simpler models lack. However, the lower recall value indicates that it is missing a large portion of the negative instances when predicting which the simpler models capture better.

4.2. Error Analysis

By calculating and analyzing the log probabilities of each predicted token by BERT and GPT, one can get a sense of how confident the models are in their prediction. This can be particularly helpful when trying to understand why the model made a particular decision or in cases where it's important to quantify the uncertainty of the model's predictions. The following review is an interesting instance:

"great flat in a cute neighbourhood close to tube and transportation we felt safe and this place could easily accommodate people or children communication for booking and check in was great however we needed assistance and help afterwards to make the oven work but didn't get any response great price for what this flat has to offer"

BERT predicted a probability of 0.7525 for a positive sentiment, while GPT was confident with 100% certainty that the review should be positive. However, the actual label of the review was negative, which may not be immediately apparent even to a human analyzing the review. This observation holds true when examining other misclassified reviews as well. In more extreme instances, such as the review stating, "*clean and good host*" which is labeled as a negative sentiment, one could argue that the review's message and the corresponding star rating for the apartment do not align whatsoever, or that it is mislabeled. These findings highlight the challenges of sentiment analysis and explain imperfection of GPT even for this simple binary classification task.

5. Discussion

5.1. Sentiment Analysis

Contrary to the hypothesis, Logistic Regression and NB demonstrated similar, or even better, performance than the fine-tuned BERT model. This could be caused by the dataset's size or complexity. As the BERT model is optimal when dealing with large, complex datasets, it could be possible that it did not use its full potential. To the contrary, NB assumes that the features are conditionally independent and thus cannot capture complex dependencies in language. Moreover, the task's simplicity could be another contributing factor. Binary sentiment analysis

might not necessitate the complex representations that BERT can provide. Furthermore, the imbalanced dataset could hinder BERT's learning capabilities. Similarly, the relatively short reviews in the dataset might reduce the performance as well, as BERT is not trained on single words or bigrams, but full sentences.

Further, as anticipated, GPT-3.5 outperformed all other models substantially in detecting negative sentiment. This could be caused by GPT-3.5 175 billion parameters, which significantly outnumber other models in terms of complexity. The combination of the level of complexity, with the model's hosting on OpenAI's servers and extensive human-supervised finetuning, might result in higher performance.

Thus, the findings suggest that the selection of the model should align with the task's characteristics and available data. While models like BERT and GPT-3.5 exemplify state-of-the-art performance for numerous NLP tasks at the same time, there will always be scenarios where simpler models, such as Logistic Regression and NB, could deliver comparable or even superior results for a specific task. This better performance on smaller tasks also comes with the benefit of reduced resource requirements, shorter runtimes, and therefore lower costs.

5.2. Topic Modeling

Using LDA, six topics were obtained of which five of them considered meaningful. The findings align to the previous study of Cheng and Jin (2019) who identified four large topics: location, amenities, host, and recommendation. The topic 'recommendation' was not as frequently present as the others. It must be mentioned that topic titles can be subjectively influenced. Interestingly, this study also identified 'booking' as a topic which was not mentioned in Cheng and Jin's (2019) paper.
Contrary to LDA, 'recommendation' appeared most frequently using Top2Vec. Using topic modeling, 'price' could not be identified as a large topic, aligning with the findings of Cheng and Jin (2019), and using Top2Vec, it was assigned to 0.07% of the reviews.

In conclusion, those results show that some topics are frequently mentioned by guests and thus seem to be considered important for them when evaluating their stay. Aribnb could use this information by trying to improve their service and quality of offerings. For instance, they could increase their selection efforts or criteria regarding 'host' or 'location'. Also, Airbnb could communicate those topics to hosts so that they can optimize the offers.

5.3. Practical & Ethical Implications

Results suggest that Airbnb should consider their choice of model carefully. When OpenAI stops limiting its API access to GPT, Airbnb could leverage the tool and thereby save developer costs for other Airbnb internal analysis. However, this could have some negative effects. For instance, Airbnb's could rely on external tools and thereby may disrupt their operations and decision-making. It would strongly depend on the functionalities of an external resource.

Considering ethical implications, OpenAI openly states on their website when accessing the model, they can reuse the data for training processes. Airbnb must have clear policies in place to handle and protect (personal) data, and users should be made aware of how their data is being used. Further, large language models are trained on biased and discriminatory text and thereby can transmit those existing biases. Thus, Airbnb should consider the outputs carefully, especially to avoid unfair treatment or discrimination against hosts or guests. Similarly, Airbnb might have difficulties communicating results to their customers when not knowing the detailed functioning of pretrained models. Also, it is important to provide clear explanations and justifications to hosts and guests regarding how the models influenced their decision-making.

7. Conclusion & Limitation

The paper aimed to develop and compare different models for binary sentiment analysis based on scraped Airbnb review comments. To get a general understanding of topics important to customers, topic modeling using LDA and Top2Vec were utilized, identifying location, amenities, host, recommendation, and booking as meaningful topics.
Based on existing literature, the hypothesis was derived that LLMs like BERT and GPT would outperform traditional models such as Logistic Regression and Naïve Bayes due to their higher level of understanding of linguistic complexities and ability to capture context and subtext. As expected GPT-3.5 demonstrated the best performance with high precision and recall. However, both Logistic Regression and Naïve Bayes performed similar or better than the BERT model. In conclusion this study suggest that the choice of model should depend on the available data and concrete task. When simpler models can accomplish similar results for less resource-intensive tasks, their application becomes valuable and more economical. These simpler models are also easier to maintain and adapt to future changes in the data.

However, similar limitations of the present study must be addressed. First, as mentioned earlier the dataset is very imbalanced containing only 1% negative reviews. This problem was also previously identified by other scholars (Alsudais & Teubner, 2019; Bridges & Vásquez, 2018).

Even though oversampling was applied to minimize the negative effects, models could have decreased performance as the variety of negative reviews and the vocabulary is too small.
Also, undersampling techniques are generally an approach to handle imbalanced classes. However, when trying undersampling the models' performances dropped. This could be due to the loss of word and context variety. BERT that leverages contextual information and sequential dependencies might be suffer by this. Logistic regression on the other hand might be negatively affected as the quality of the coefficient estimates gets reduced.
Future research could combine datasets from different sources and thus generate larger negative comment diversity.
Closely related to this, as mentioned in the discussion short comments could have negative consequences. Particularly the BERT model could lose valuable context information. Therefore, the focus could be on longer reviews in the future.

It is possible that the BERT model was not a good fit for the task at hand. The initial BERT model was trained to perform language understanding tasks and predicting the next sentence based on given input. Therefore, the features extracted by the DistilBERT model might not be the best for determining sentiment in the relatively short input data that we have. This might also be the reason that even relatively complex classifier for a binary classification does not lead to overfitting on the binary classification.

Furthermore, exploring the potential of training the whole BERT model, rather than just a classifier on top, could provide insights into the benefits of this approach, particularly for more complex classification tasks, despite its computational demands. However, choosing another NLP model with a stronger focus on individual tokens might result in better performance.

Considering GPT-3.5 it must be highlighted that its performance was only evaluated based on a subset of the test data due to the API costs. Thus, comparability with other models might be restricted. Finally, future research could examine the use of GPT-3.5 and its performance on a larger dataset. Also, it would be interesting to apply advanced models like GPT-3.5 in more challenging tasks. Identifying the domains where their human-like capabilities can be most effectively utilized would guide the practical application and give insights into their full potential.

References

Airbnb. (2022). Airbnb Newsroom. https://news.airbnb.com/about-us/

Alsudais, A., & Teubner, T. (2019). *Large-scale sentiment analysis on airbnb reviews from 15 cities*. 25th Americas Conference on Information Systems.

Angelov, D. (2020). *Top2Vec: Distributed Representations of Topics*. https://doi.org/10.48550/ARXIV.2008.09470

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Bridges, J., & Vásquez, C. (2018). If nearly all *Airbnb* reviews are positive, does that make them meaningless? *Current Issues in Tourism*, *21*(18), 2065–2083. https://doi.org/10.1080/13683500.2016.1267113

Brooks. (1957). *"Word-of-mouth" advertising in selling new products*. https://journals.sagepub.com/doi/pdf/10.1177/002224295702200205?casa_token=Umb4rTCNjWsAAAAA:FShP-5-nJDLuymHPU8qnBFIGQca12aaoNTg_g0OS1JggzQy7vVPomWCaLovvr3qfHj9wDoOLHEBCQJg

Cheng, M., & Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management*, *76*, 58–70. https://doi.org/10.1016/j.ijhm.2018.04.004

Chiny, M., Bencharef, O., Hadi, M. Y., & Chihab, Y. (2021). A Client-Centric Evaluation System to Evaluate Guest's Satisfaction on Airbnb Using Machine Learning and NLP. *Applied Computational Intelligence and Soft Computing*, *2021*, 1–14. https://doi.org/10.1155/2021/6675790

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805

Feldges, C. (2022, November 9). Text Classification with BERT in TensorFlow and PyTorch. *Medium*. https://medium.com/@claude.feldges/text-classification-with-bert-in-tensorflow-and-pytorch-4e43e79673b3

Glazkova, A. (2020). *A Comparison of Synthetic Oversampling Methods for Multi-class Text Classification* (arXiv:2008.04636). arXiv. http://arxiv.org/abs/2008.04636

Guttentag, D. (2019). Progress on Airbnb: A literature review. *Journal of Hospitality and Tourism Technology*, *10*(4), 814–844. https://doi.org/10.1108/JHTT-08-2018-0075

Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the Knowledge in a Neural Network* (arXiv:1503.02531). arXiv. http://arxiv.org/abs/1503.02531

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed). Pearson Prentice Hall.

Khomsah, S. (2020). Naive Bayes Classifier Optimization on Sentiment Analysis of Hotel Reviews. *Jurnal Penelitian Pos Dan Informatika*, *10*(2), 157. https://doi.org/10.17933/jppi.2020.100206

Kulkarni, A., Chong, D., & Batarseh, F. A. (2021). *Foundations of data imbalance and solutions for a data democracy*. https://doi.org/10.48550/ARXIV.2108.00071

Lawani, A., Reed, M. R., Mark, T., & Zheng, Y. (2019). Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston. *Regional Science and Urban Economics*, *75*, 22–34. https://doi.org/10.1016/j.regsciurbeco.2018.11.003

Lee, C. K. H., Tse, Y. K., Zhang, M., & Ma, J. (2020). Analysing online reviews to investigate customer behaviour in the sharing economy: The case of Airbnb. *Information Technology & People*, *33*(3), 945–961. https://doi.org/10.1108/ITP-10-2018-0475

Liu, Y. (2006). *Word of mouth for movies: Its dynamics and impact on box office revenue*. https://journals.sagepub.com/doi/pdf/10.1509/jmkg.70.3.074?casa_token=04LpgTuZ_wEAAAAA:0g7Dy7v8j_-kxf-yOiYHwwSIloCcpL9HULRZpMahLgFrewwaP0FCv7xyL1ez9AuNUEqGtDu74GnWyr0

Luo, Y., & Tang, R. (Liang). (2019). Understanding hidden dimensions in textual reviews on Airbnb: An application of modified latent aspect rating analysis (LARA). *International Journal of Hospitality Management*, *80*, 144–154. https://doi.org/10.1016/j.ijhm.2019.02.008

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Mellinas, J. P., & Reino, S. (2019). Average scores integration in official star rating scheme. *Journal of Hospitality and Tourism Technology*, *10*(3), 339–350. https://doi.org/10.1108/JHTT-07-2017-0050

OpenAI API. (2023). https://platform.openai.com

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations* (arXiv:1802.05365). arXiv. http://arxiv.org/abs/1802.05365

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*.

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Raza, M. R., Hussain, W., & Varol, A. (2022). Performance Analysis of Deep Approaches on Airbnb Sentiment Reviews. *2022 10th International Symposium on Digital Forensics and Security (ISDFS)*, 1–5. https://doi.org/10.1109/ISDFS55398.2022.9800816

Rezazadeh, P. K., Nikolenko, L., & Rezaei, H. (2021). Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), *Machine Learning and Knowledge Extraction* (pp. 173–184). Springer International Publishing. https://doi.org/10.1007/978-3-030-84060-0_11

Rokou, T. (2022, June 13). Airbnb: These are Europe's most profitable cities. *TravelDailyNews International*. https://www.traveldailynews.com/hotels-lodging/airbnb-these-are-europes-most-profitable-cities/

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (arXiv:1910.01108). arXiv. http://arxiv.org/abs/1910.01108

Santos, G., Mota, V. F. S., Benevenuto, F., & Silva, T. H. (2020). Neutrality may matter: Sentiment analysis in reviews of Airbnb, Booking, and Couchsurfing in Brazil and USA. *Social Network Analysis and Mining*, *10*(1), 45. https://doi.org/10.1007/s13278-020-00656-5

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, *30*(4), 415–433.

Von Hoffen, M., Hagge, M., Betzing, J. H., & Chasin, F. (2018). Leveraging social media to gain insights into service delivery: A study on Airbnb. *Information Systems and E-Business Management*, *16*(2), 247–269. https://doi.org/10.1007/s10257-017-0358-7

Wang, Z., Xie, Q., Ding, Z., Feng, Y., & Xia, R. (2023). *Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study* (arXiv:2304.04339). arXiv. http://arxiv.org/abs/2304.04339

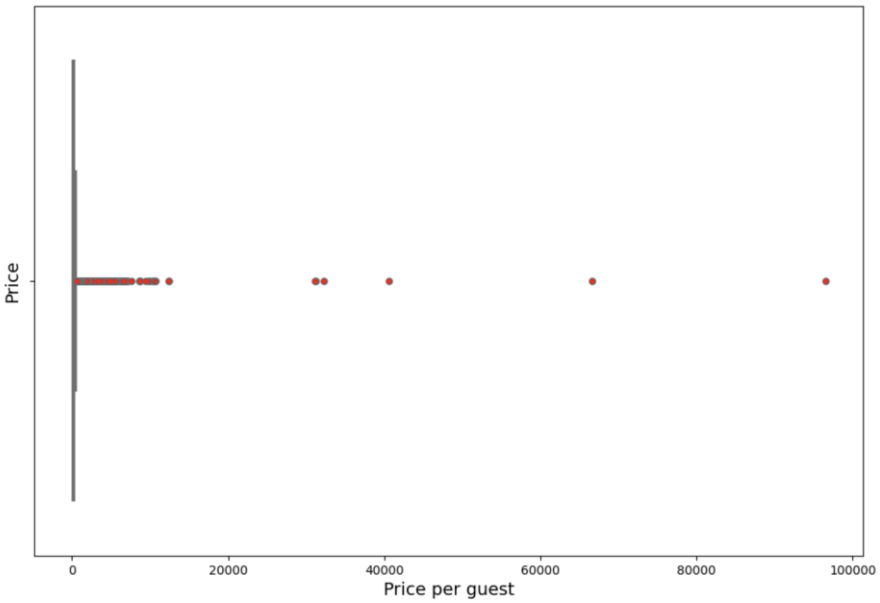
Weyerer, J. C. (2019). *Online Review Credibility*. *20*(1).

Zhong, Q., Ding, L., Liu, J., Du, B., & Tao, D. (2023). *Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT* (arXiv:2302.10198). arXiv. http://arxiv.org/abs/2302.10198

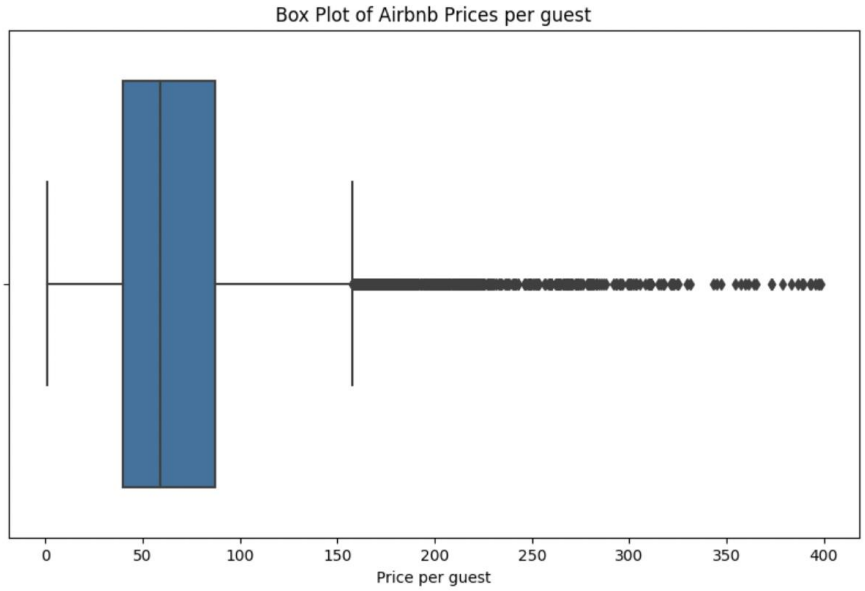
Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). *Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books*. https://doi.org/10.48550/ARXIV.1506.06724

Appendix

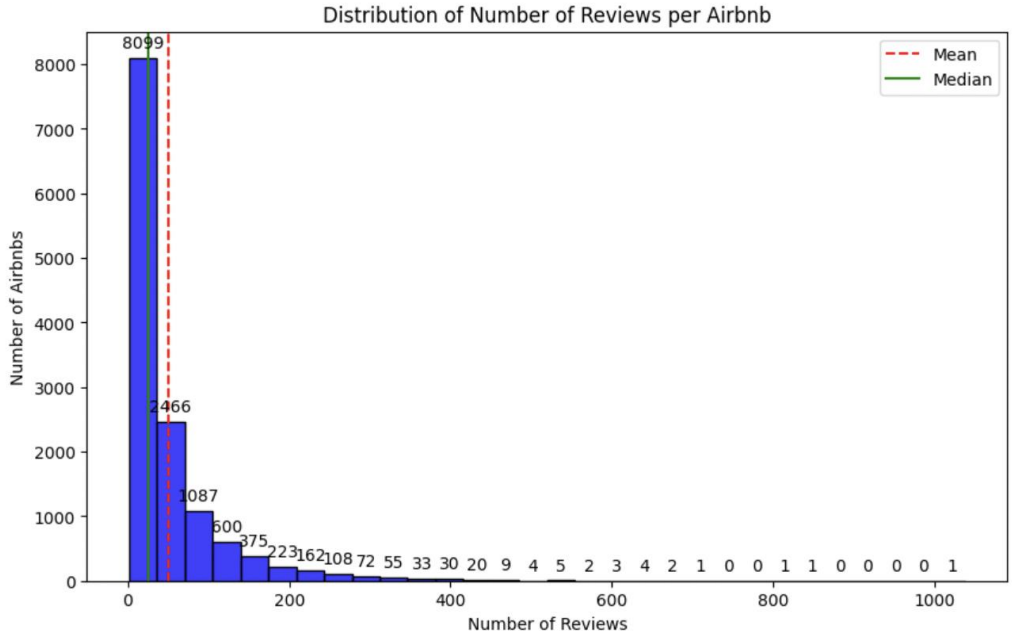
Boxplot Price per Guest (Before Filtering)



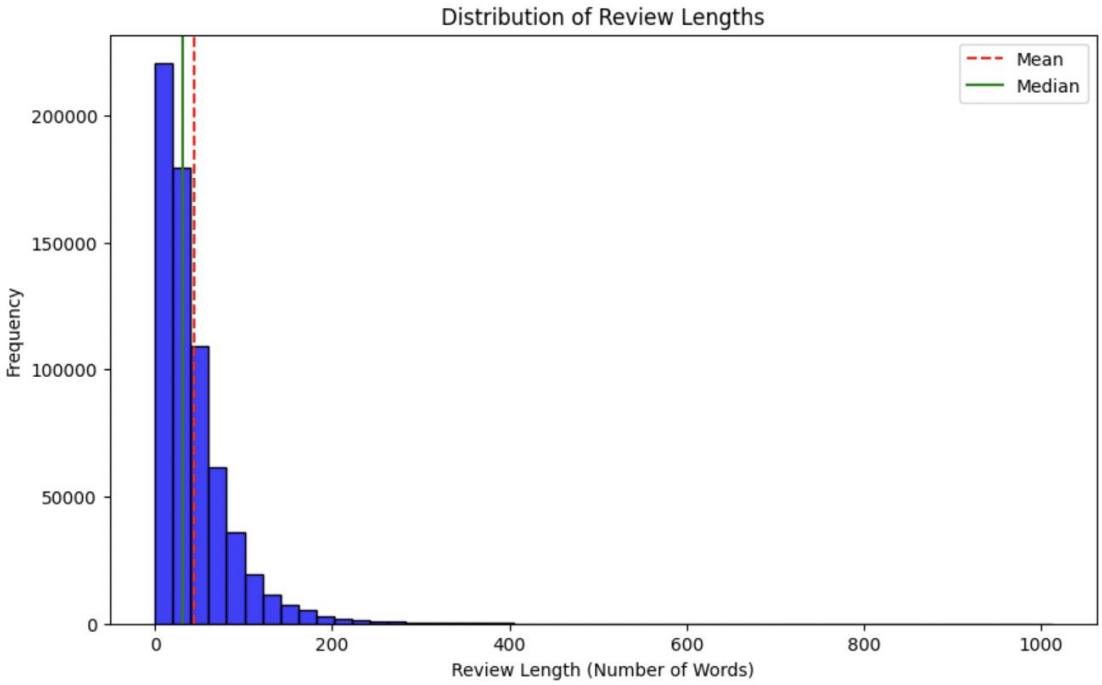
Boxplot Price per Guest (After Filtering)



Distribution of Number of Reviews per Airbnb



Distribution of Review Lengths



Boxplot Superhosts versus Non-Superhosts