Aspect Extraction for Hotel Reviews: A Hybrid Approach

Darya Solomenko - 5 August, 2024

Aspect-based sentiment analysis

Unlike traditional sentence- or document-level sentiment analysis, aspect-based sentiment analysis directly identifies and evaluates opinions, comprising both a sentiment polarity (positive or negative) and a corresponding opinion target.

> **"The bed was comfy, our main issue was just that we got repeatedly woken up all night by a dog barking very close by and was allowed to continue barking all night and in the early hours of the morning. Very frustrating but if that hadnt been a problem, then no issues with the accommodation."**

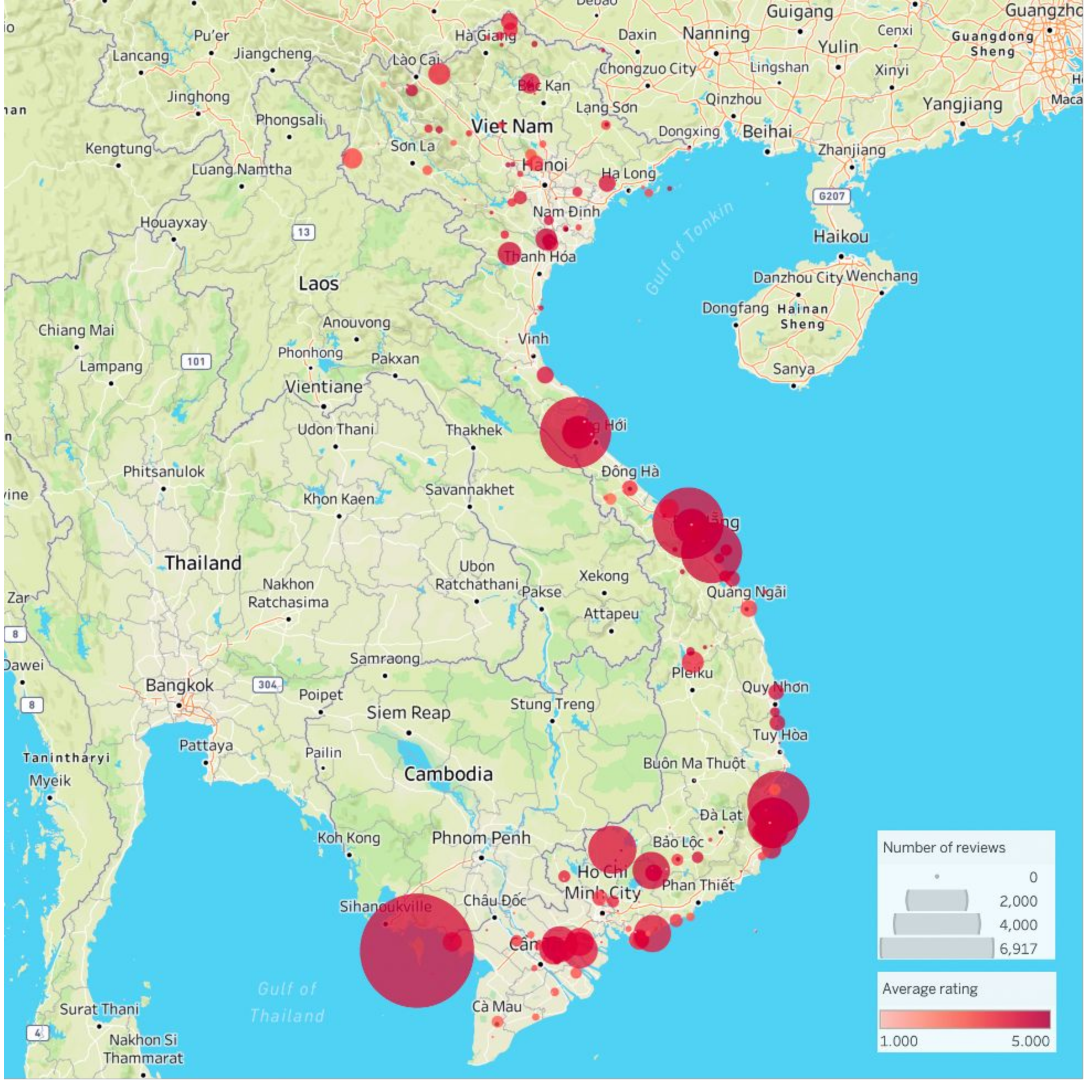
– A guest review on TripAdvisor

Data set

32,449 English language customer reviews of Vietnamese hotels between 2005 and 2022 from the TripAdvisor website.

Additional data sets:

- Geo locations
- Accommodation categories

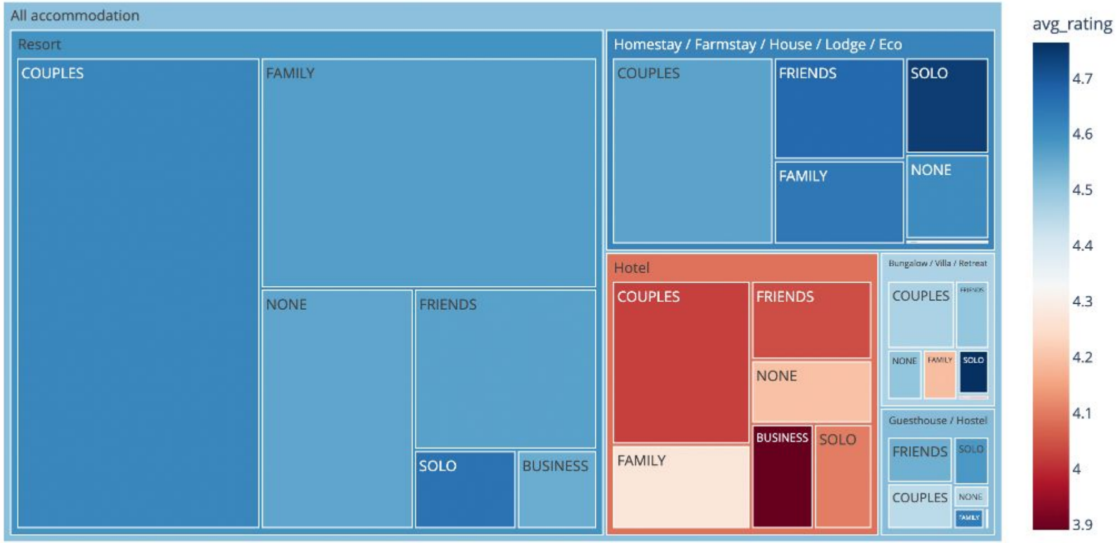


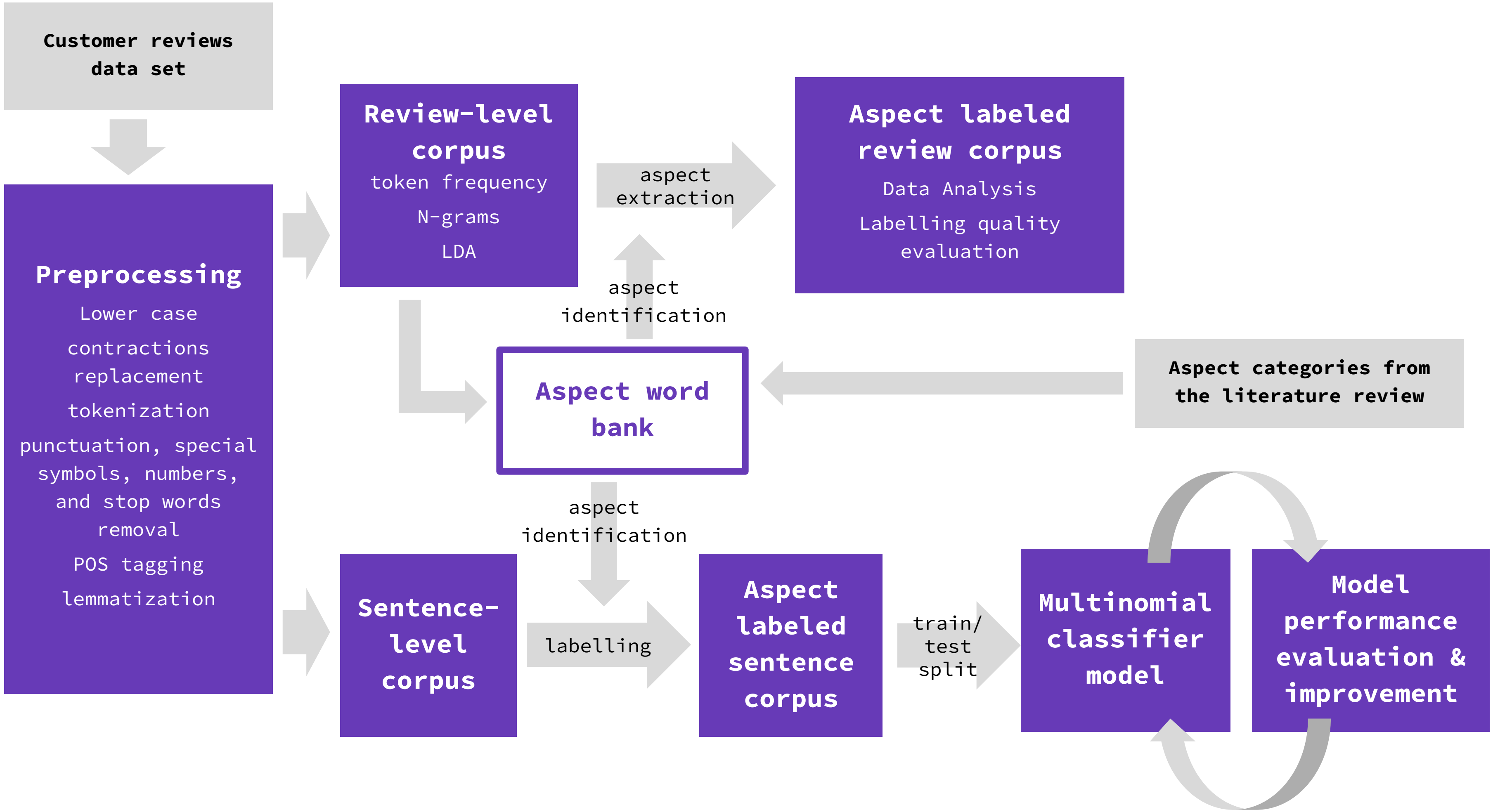
Exploration of data

Couples staying in resorts - the largest group of reviews.

Hotels receive lower user ratings while family-owned accommodation (homestays, lodges etc.) - higher.

Resorts display the longest average review length (152.44 words). The average review for a guesthouse or a hostel is 30% shorter.





Customer reviews data set

Preprocessing

Lower case

contractions replacement

tokenization

punctuation, special symbols, numbers, and stop words removal

POS tagging

lemmatization

Review-level corpus

token frequency

N-grams

LDA

aspect extraction

Aspect labeled review corpus

Data Analysis

Labelling quality evaluation

aspect identification

Aspect word bank

Aspect categories from the literature review

aspect identification

Sentence-level corpus

labelling

Aspect labeled sentence corpus

train/ test split

Multinomial classifier model

Model performance evaluation & improvement

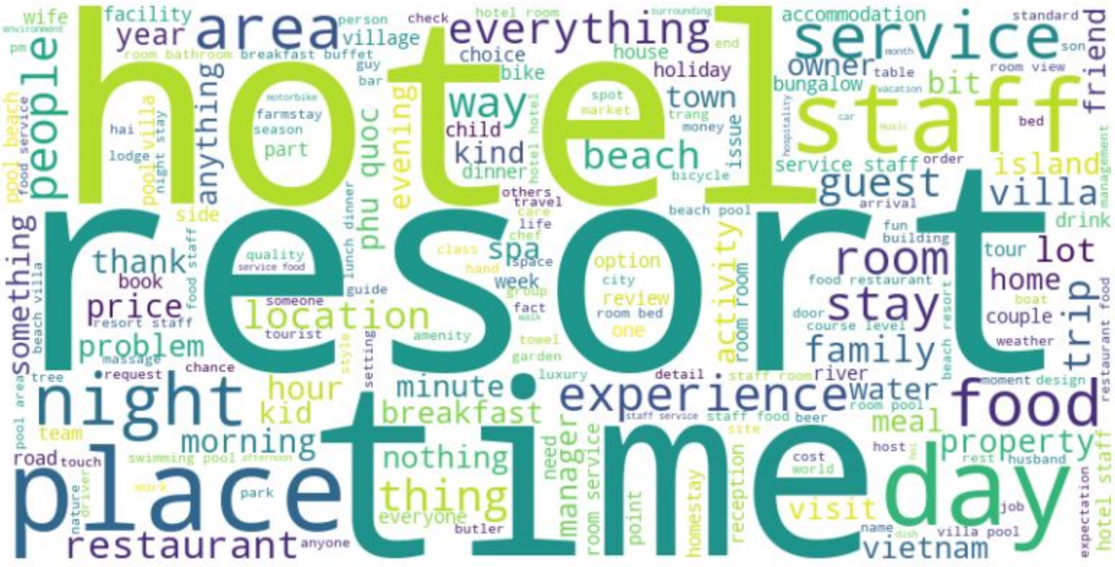
Aspect Identification

Seven aspects: location, cleanliness, comfort, facilities, service, value, food and beverage.

A vocabulary-based approach was applied for aspect identification and extraction.

Each review might include 0, 1 or more mentions of each seven aspects.

On average, each review mentions 4.60 ± 0.01 (95% CI) aspects.

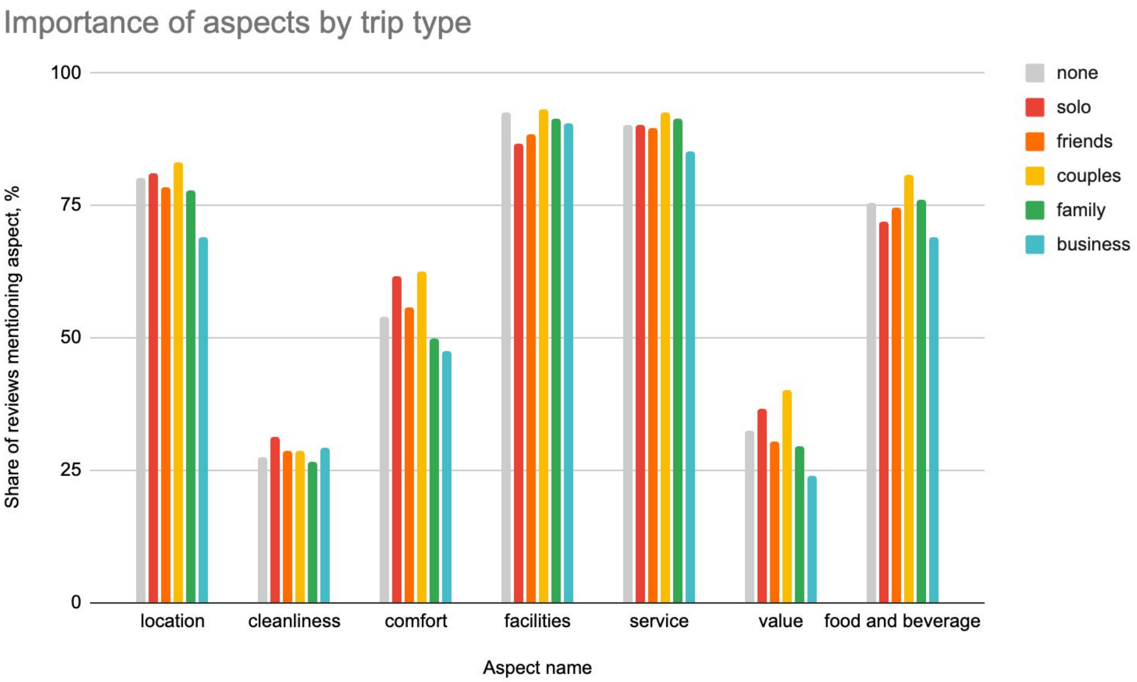


What aspects matter?

While the length of reviews varies greatly based on the trip type, average number of aspects per review stays between 4.14 and 4.81.

No difference between accommodation categories in the number of aspects mentioned per review.

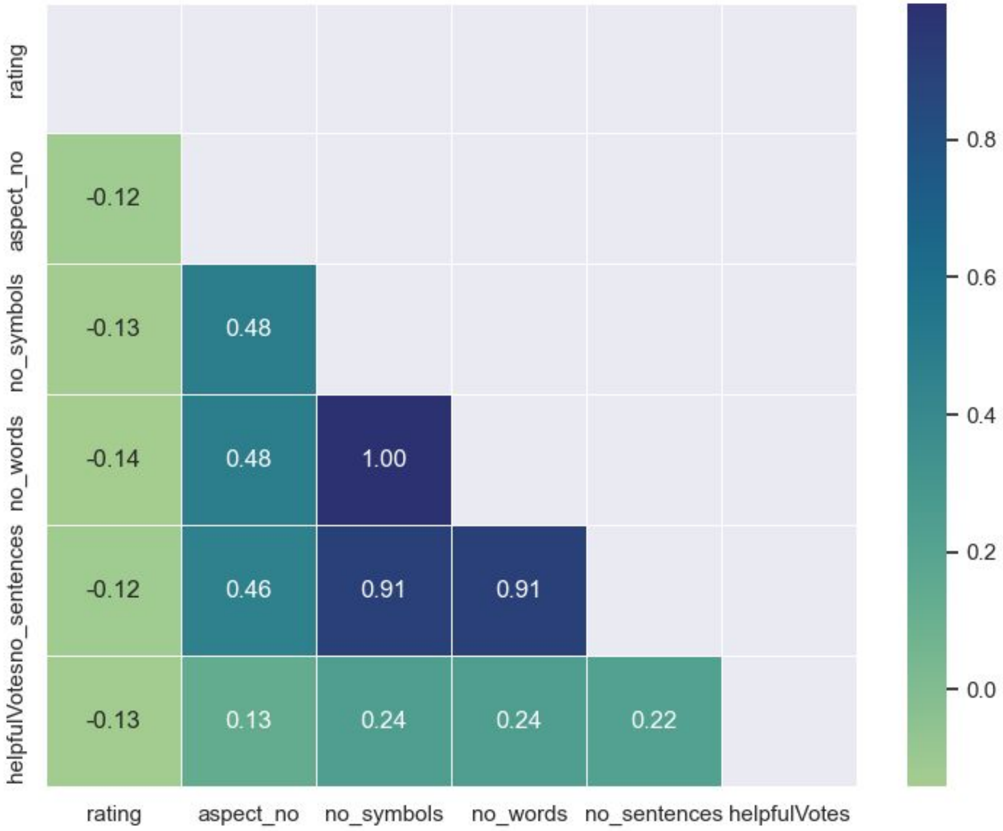
5-star reviews have the lowest average number of aspects mentioned, 2 and 3-star the highest.



Aspects vs Ratings

Moderate correlation between the review length and the number of aspects in it.

No correlation between the number of aspects mentioned in a review and a rating given by a user.



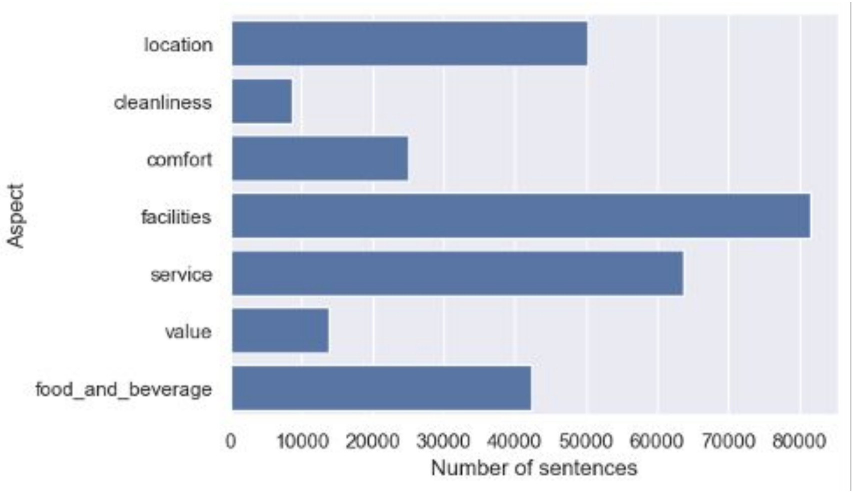
Aspect Classification

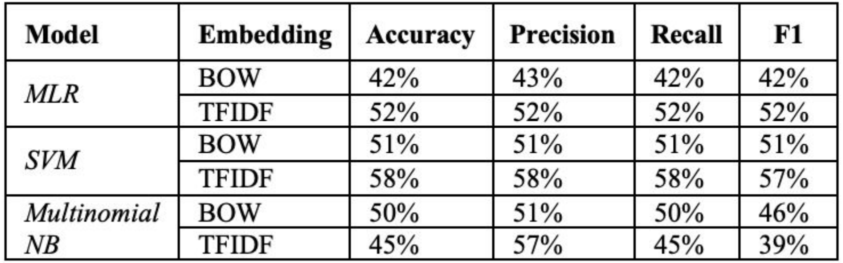
3 models: Logistic Regression, Stochastic Gradient Descent classifier (with SVM loss parameter) and Multinomial Naïve Bayes.

Unbalanced dataset: 10x different classes.

Measures taken:

- stratified sampling
- 5-fold cross-validation for hyperparameter tuning
- balanced weight class parameter
- random undersampling



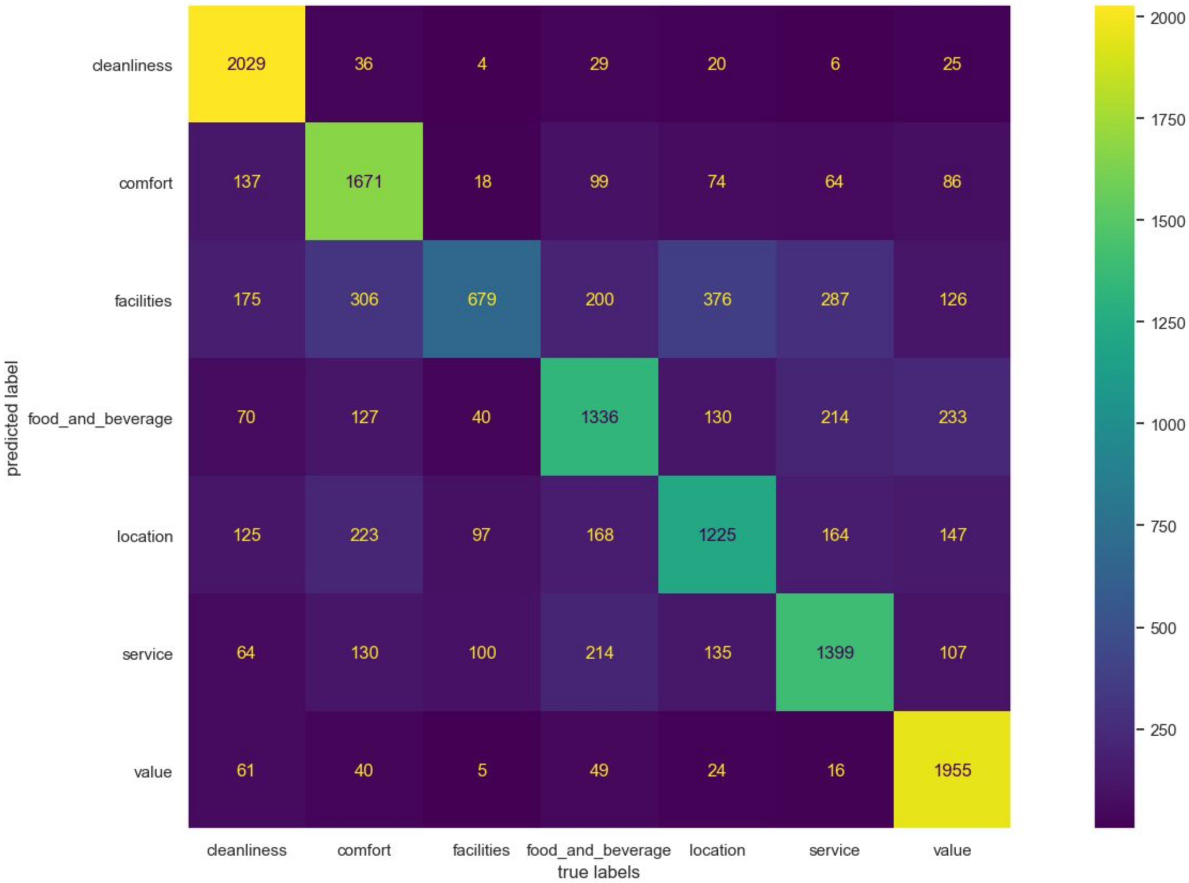


Model	Embedding	Accuracy	Precision	Recall	F1
MLR	BOW	42%	43%	42%	42%
	TFIDF	52%	52%	52%	52%
SVM	BOW	51%	51%	51%	51%
	TFIDF	58%	58%	58%	57%
Multinomial NB	BOW	50%	51%	50%	46%
	TFIDF	45%	57%	45%	39%

Balancing data set

The best performance on the balanced data set was achieved for the **SVM BOW classifier** model with a significant increase across all performance metrics:

- accuracy to 68%
- precision 68%
- recall 68%
- F1 score 67%



	cleanliness	comfort	facilities	food_and_beverage	location	service	value
cleanliness	2029	36	4	29	20	6	25
comfort	137	1671	18	99	74	64	86
facilities	175	306	679	200	376	287	126
food_and_beverage	70	127	40	1336	130	214	233
location	125	223	97	168	1225	164	147
service	64	130	100	214	135	1399	107
value	61	40	5	49	24	16	1955

Conclusion

- The identification of aspects uncovered **valuable insights into aspect popularity** across guest types and accommodation categories.
- The strong performance of the vocabulary-based aspect identification approach combined with the classification model for aspect class prediction **show promise for practical application on unlabelled data sets** in the hospitality domain.

Future research:

- further **automation and accuracy improvement**
- integrating this methodology with **hotel management processes**
- expanding to **more countries and and languages**

Thank you! Questions?