Model Performance

Model Name: LocationClassificationV15 **Test Date:** 21/03/2022 15:29:47 **Creator:** Giovanni Triulzi

Overview

ML Principle:
Linear Discriminant Analysis

References:

- LDA Doc.
- Stanford NLP Course
- Stanford NLP Lecture
- Engilsh Stopwords

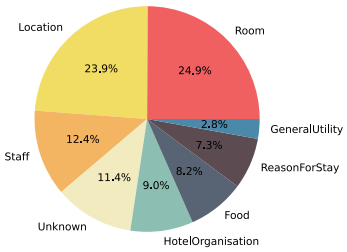
Algorithm Description:
The learning algorithm used in this classification is Linear Discriminant Analysis. This approach was chosen as it is easy to implement and is computational very efficient. The first step in the classification pipeline is removing all stop words for example 'i', 'me', etc. A list of English stop words is provided by the nltk module. Next the sentence is passed through a stemmer and a lemmatizer. Stemming just removes or stems the last few characters of a word, often leading to incorrect meanings and spelling. Lemmatization considers the context and converts the word to its meaningful base form, which is called Lemma. This is done with the SnowBallStemmer and WordNetLemmatizer class from the nltk module. The final preprocessing step is to vectorize the sentence. For this the Tf-idf vectorizer from sklearn is used. If a Tf-idf vectorizer is used the sentences don't have to be tokenized. The sentence is now represented in a numerical feature vector which now can be passed to the LDA classifier.

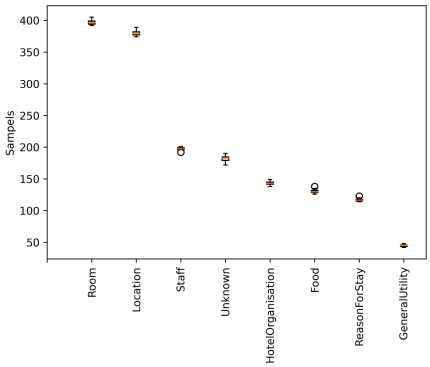
Metrics

Data: ClassifiedDataSetV1.2 with 10 folds cross validation
Split seed: 3.390625

Training Dataset

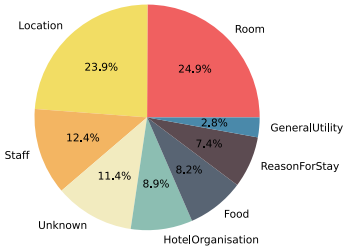
Classes	Number of samples
Room	397
Location	380
Staff	197
Unknown	181
HotelOrganisation	143
Food	130
ReasonForStay	117
GeneralUtility	45

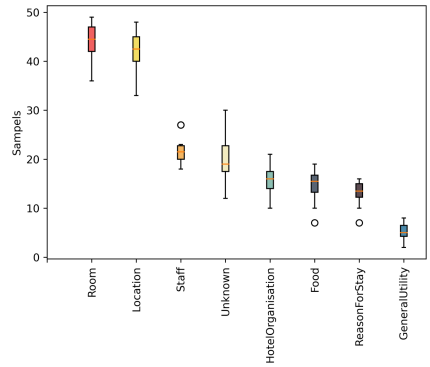


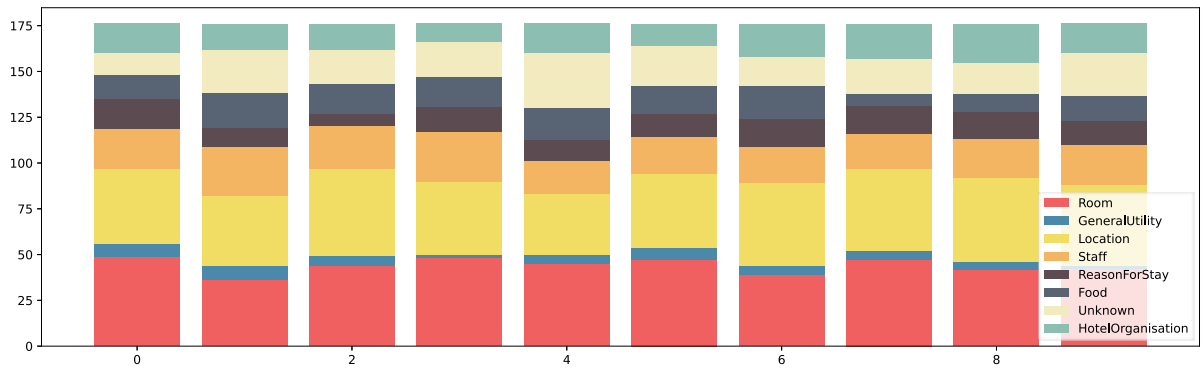


Test Dataset

Classes	Number of samples
Room	43
Location	42
Staff	21
Unknown	20
HotelOrganisation	15
Food	14
ReasonForStay	13
GeneralUtility	5

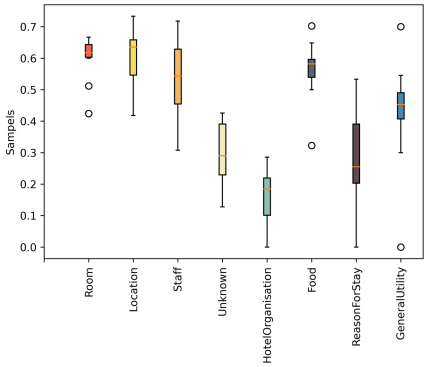




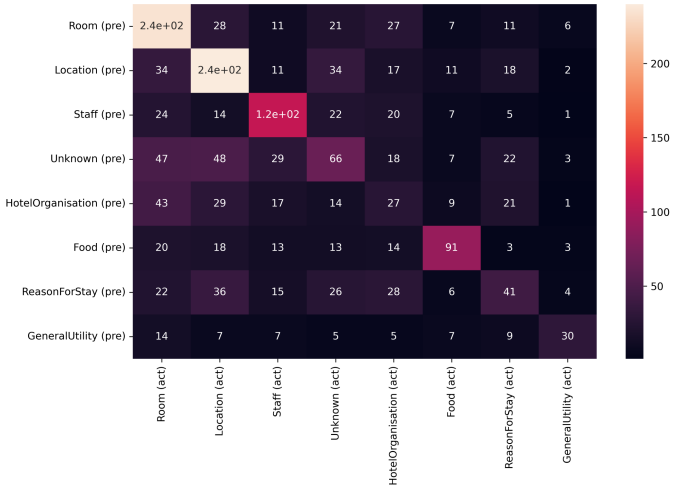


Classification Performance

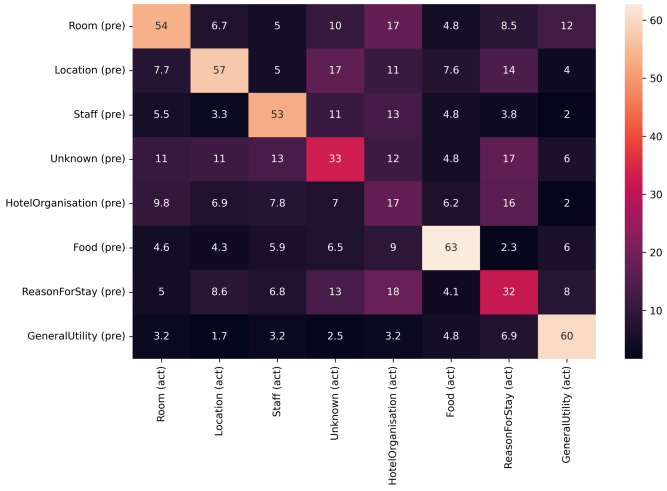
Classes	Precision	Recall	F1 Score
Room	67.92%	53.53%	59.87%
Location	65.40%	57.14%	60.99%
Staff	55.50%	52.97%	54.21%
Unknown	27.50%	32.84%	29.93%
HotelOrganisation	16.77%	17.31%	17.03%
Food	52.00%	62.76%	56.88%
ReasonForStay	23.03%	31.54%	26.62%
GeneralUtility	35.71%	60.00%	44.78%
Accuracy			48.07%
Macro Average	42.98%	46.01%	43.79%
Weighted Average	51.08%	48.07%	49.09%



ConfusionMatrix:



Normalised ConfusionMatrix:



F1 Socre by split: