Model Performance

Model Name: LocationClassificationV2 **Test Date:** 21/03/2022 14:06:12 **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.

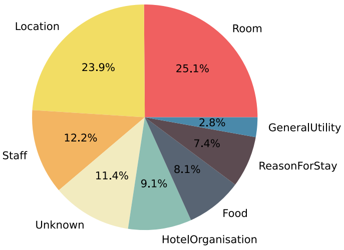
Metrics

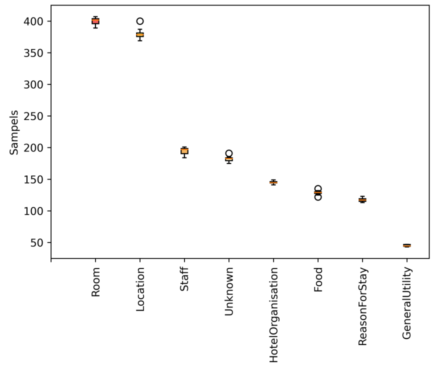
Data: ClassifiedDataSetV1.2 with 10 folds cross validation

Split seed: 3.15625

Training Dataset

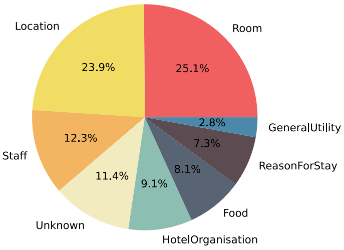
Classes	Number of samples
Room	399
Location	380
Staff	194
Unknown	181
HotelOrganisation	144
Food	128
ReasonForStay	117
GeneralUtility	45

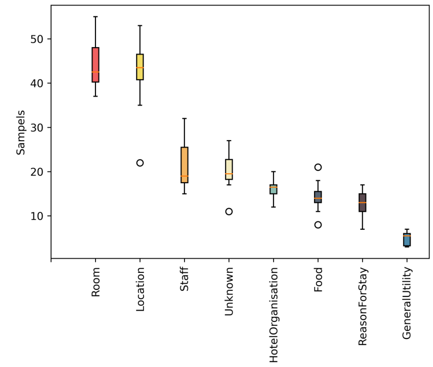


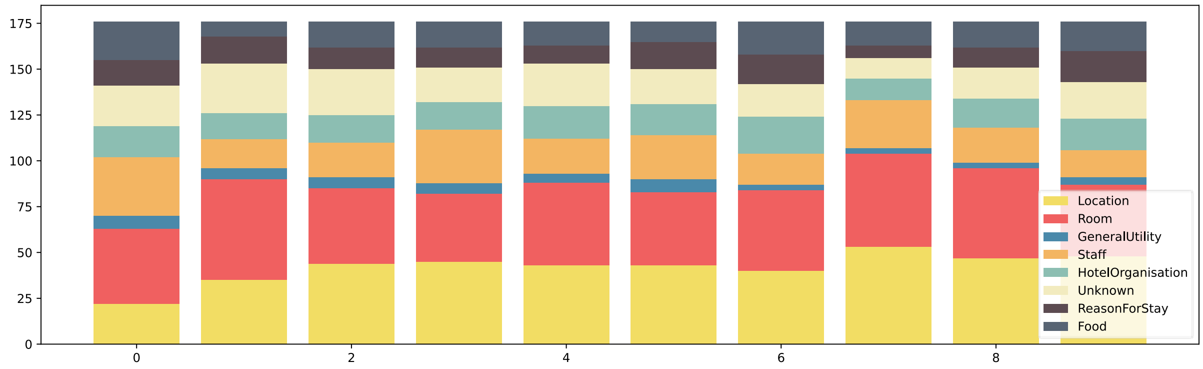


Test Dataset

Classes	Number of samples
Room	44
Location	42
Staff	21
Unknown	20
HotelOrganisation	16
Food	14
ReasonForStay	12
GeneralUtility	5

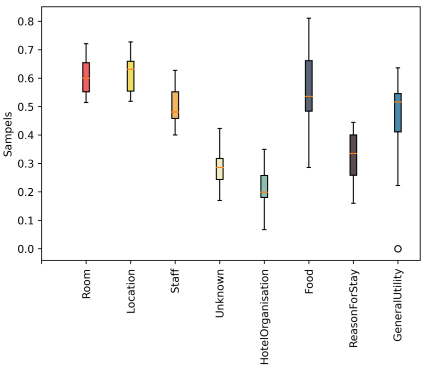




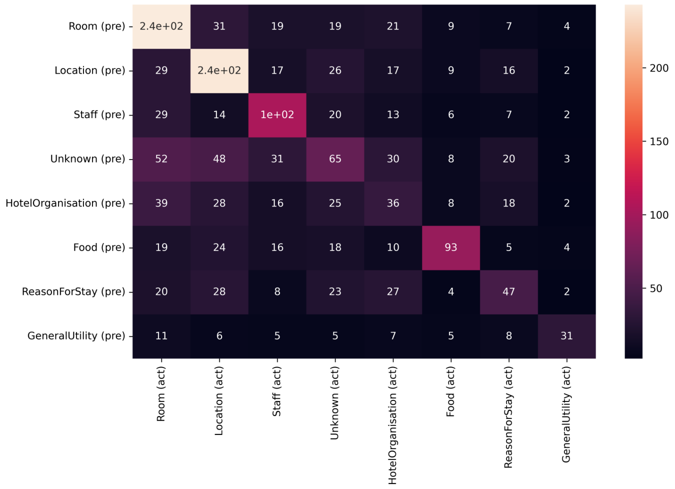


Classification Performance

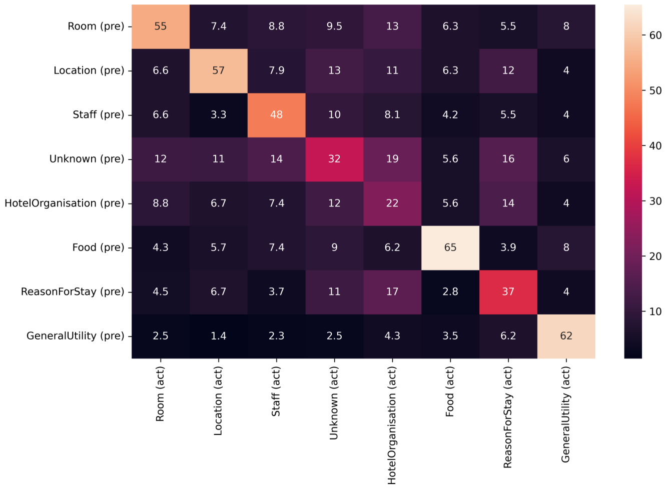
Classes	Precision	Recall	F1 Score
Room	68.84%	54.98%	61.13%
Location	67.51%	57.38%	62.03%
Staff	53.33%	48.15%	50.61%
Unknown	25.29%	32.34%	28.38%
HotelOrganisation	20.93%	22.36%	21.62%
Food	49.21%	65.49%	56.19%
ReasonForStay	29.56%	36.72%	32.75%
GeneralUtility	39.74%	62.00%	48.44%
Accuracy			48.86%
Macro Average	44.30%	47.43%	45.15%
Weighted Average	51.99%	48.86%	49.88%



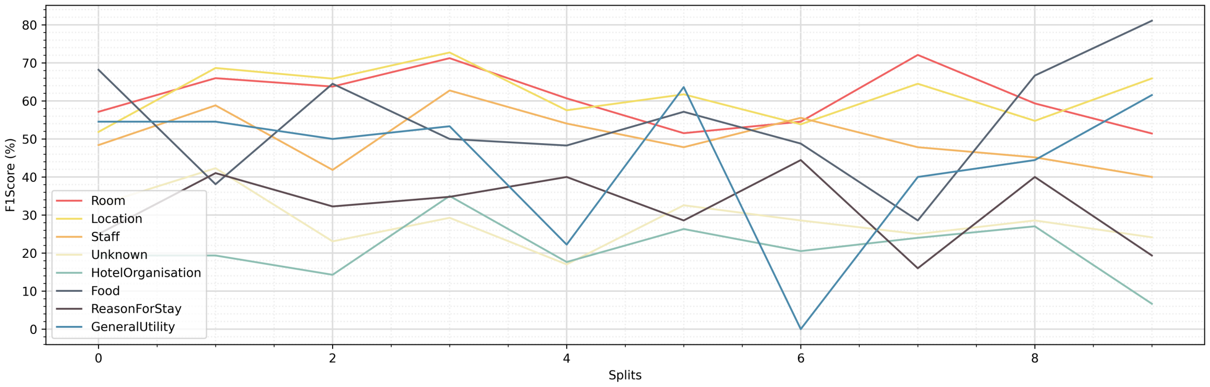
ConfusionMatrix:



Normalised ConfusionMatrix:



F1 Socre by split:


F1Score (%)
Splits
Room
Location
Staff
Unknown
HotelOrganisation
Food
ReasonForStay
GeneralUtility