Model Performance

Model Name: LocationClassificationV1 **Test Date:** 17/03/2022 19:20:25 **Creator:** Tobias Rothlin

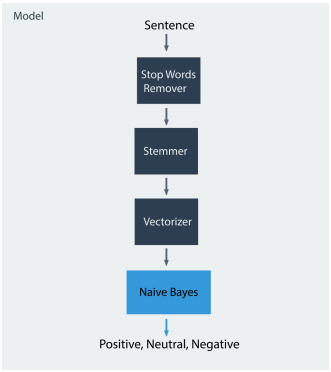
Overview

ML Principle:
Multinomial Naive Bayes

References:

- NultinomialNB Explained
- Stanford NLP Course
- Stanford NLP Lecture
- Englsh Stopwords

Algorithm Description:
The learning algorithm used in this classification is the Multinomial Naïve Bayes. 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 strop words for example 'i', 'me', 'my', 'myself', etc. A list of English stop word is provided by the nltk module. The stop words remover just removes every word that is in the list of stop words. Next the sentence is passed through the stemmer. Stemmers remove morphological affixes from words, leaving only the word stem. This is done with the PorterStemmer class from the nltk module. The final preprocessing step is to vectorize the sentence. This results in a bag of words representation of the sentence. First all the words must be tokenized and then counted. The result will be a numerical feature vector. To generate this vector the CountVectorizer class from sklearn is used. This class implements both tokenization and occurrence counting in a single class. With the sentence now represented in a vector the Naïve Bayes classifier can work with this vector. For the implementation of the Naïve Bayes classifier the MultinomialNB class (sklearn) is used.



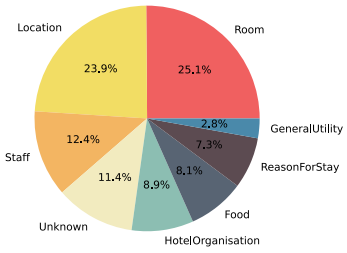
Classification Pipeline

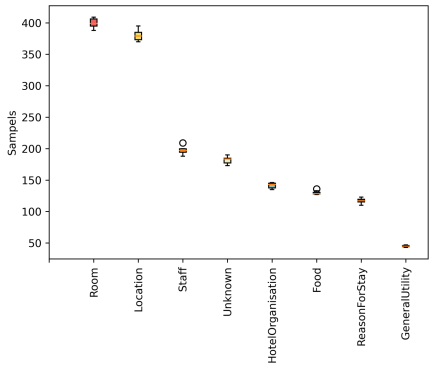
Metrics

Data: ClassifiedDataSetV1.2 with 10 folds cross validation
Split seed: 4.989823

Training Dataset

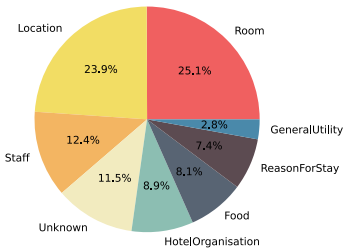
Classes	Number of samples
Room	399
Location	380
Staff	197
Unknown	181
HotelOrganisation	141
Food	129
ReasonForStay	117
GeneralUtility	45

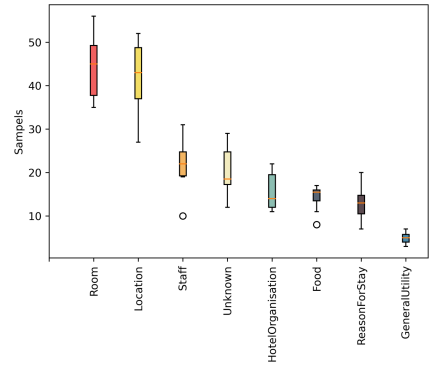


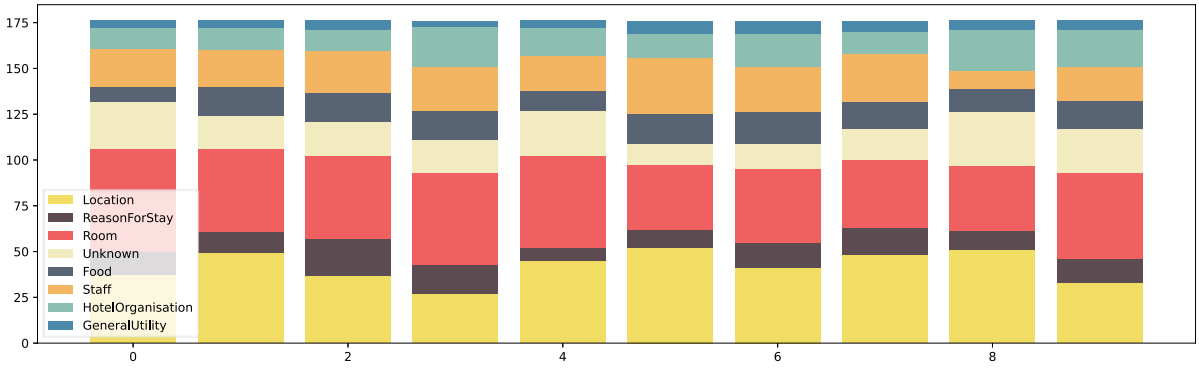


Test Dataset

Classes	Number of samples
Room	44
Location	42
Staff	21
Unknown	20
HotelOrganisation	15
Food	14
ReasonForStay	13
GeneralUtility	5

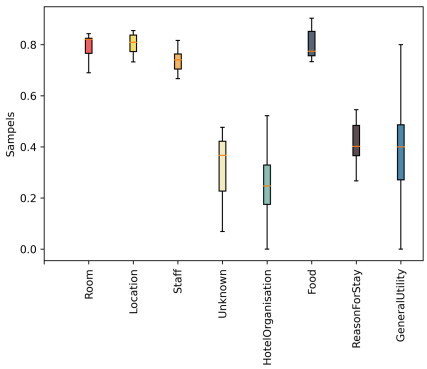




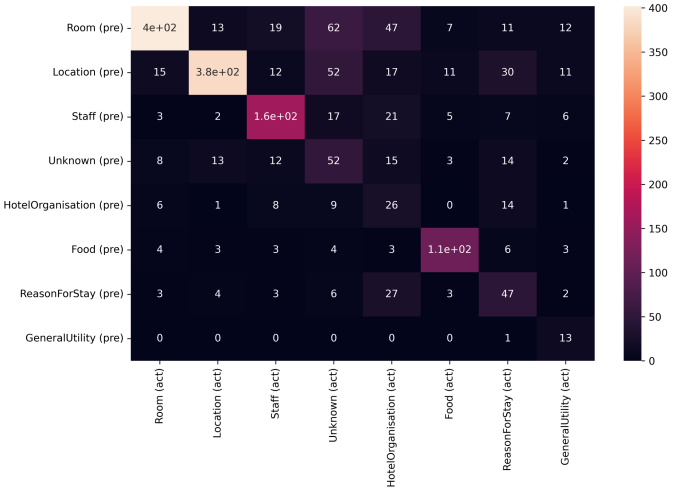


Classification Performance

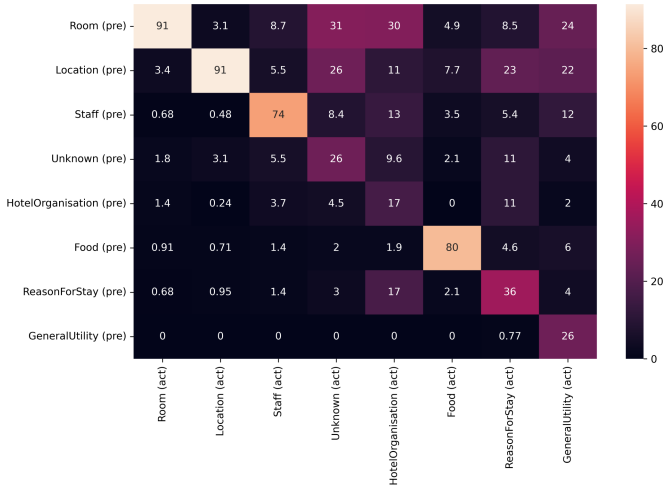
Classes	Precision	Recall	F1 Score
Room	70.16%	91.16%	79.29%
Location	72.18%	91.43%	80.67%
Staff	72.52%	73.85%	73.18%
Unknown	43.70%	25.74%	32.40%
HotelOrganisation	40.00%	16.67%	23.53%
Food	81.43%	79.72%	80.57%
ReasonForStay	49.47%	36.15%	41.78%
GeneralUtility	92.86%	26.00%	40.62%
Accuracy			68.12%
Macro Average	65.29%	55.09%	56.51%
Weighted Average	65.26%	68.12%	64.77%



ConfusionMatrix:



Normalised ConfusionMatrix:



F1 Socre by split: