Model Performance

Model Name: MultinomialNaiveBayesOnLocation **Test Date:** 23/03/2022 15:55:16 **Creator:** Tobias Rothlin

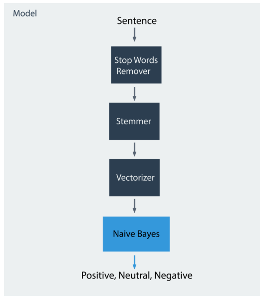
Overview

ML Principle:
Multinomial Naive Bayes

References:

- NultinomialNB Explained
- Stanford NLP Course
- Stanford NLP Lecture
- Engilsh 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.



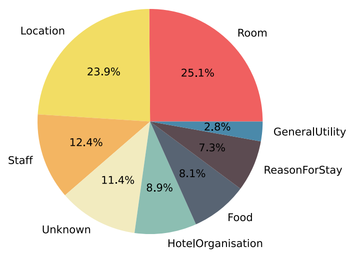
Classification Pipeline

Metrics

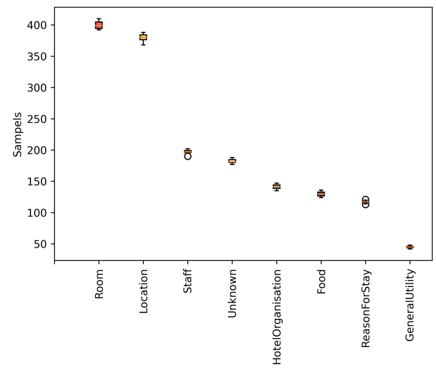
Data: ClassifiedDataSetV1.2 with 10 folds cross validation
Split seed: 4.83819

Training Dataset
(average)

Classes	Number of samples
Room	399
Location	379
Staff	197
Unknown	181
HotelOrganisation	141
Food	129
ReasonForStay	117
GeneralUtility	45



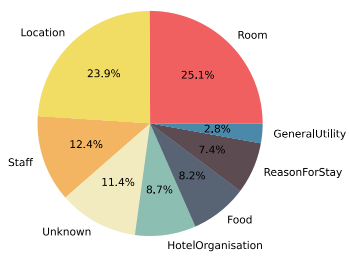
Average distribution of the samples



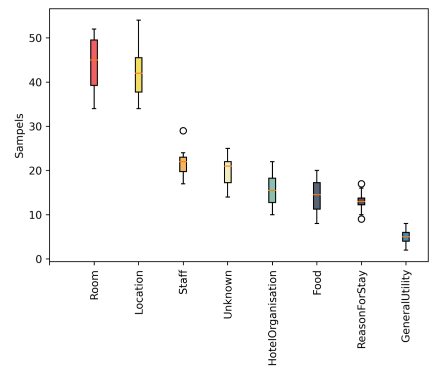
Distribution of the samples contained in each test split

Test Dataset
(average)

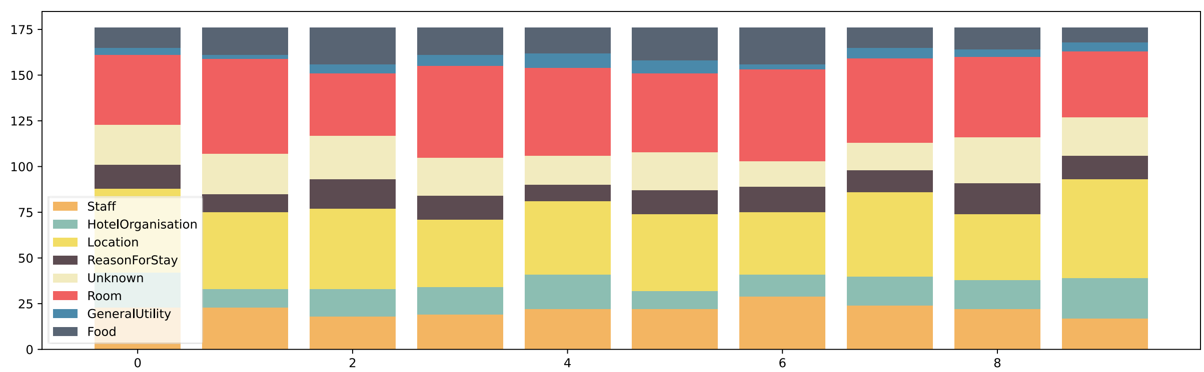
Classes	Number of samples
Room	44
Location	42
Staff	21
Unknown	20
HotelOrganisation	15
Food	14
ReasonForStay	13
GeneralUtility	5



Average distribution of the samples



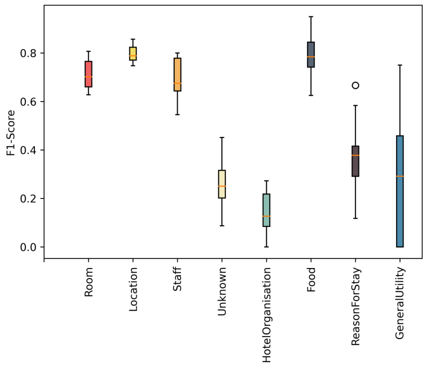
Distribution of the samples contained in each test split



Detailed training split composition

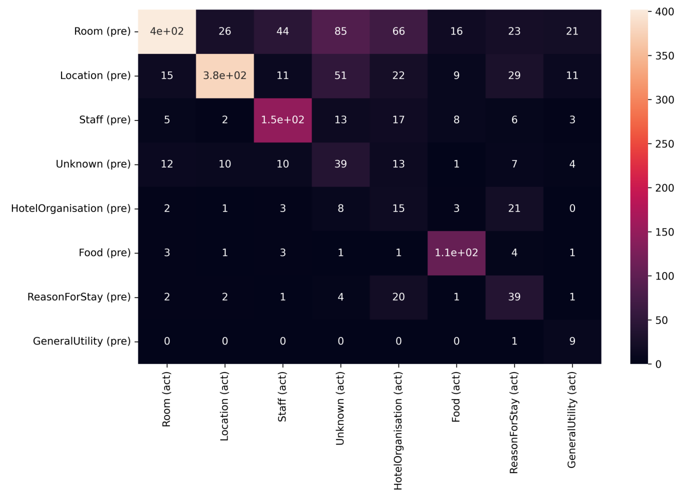
Classification Performance

Classes	Precision	Recall	F1 Score
Room	58.86%	91.16%	71.53%
Location	71.92%	90.02%	79.96%
Staff	73.13%	67.12%	70.00%
Unknown	40.62%	19.40%	26.26%
HotelOrganisation	28.30%	9.74%	14.49%
Food	88.33%	73.61%	80.30%
ReasonForStay	55.71%	30.00%	39.00%
GeneralUtility	90.00%	18.00%	30.00%
Accuracy			64.55%
Macro Average	63.36%	49.88%	51.44%
Weighted Average	62.07%	64.55%	60.33%

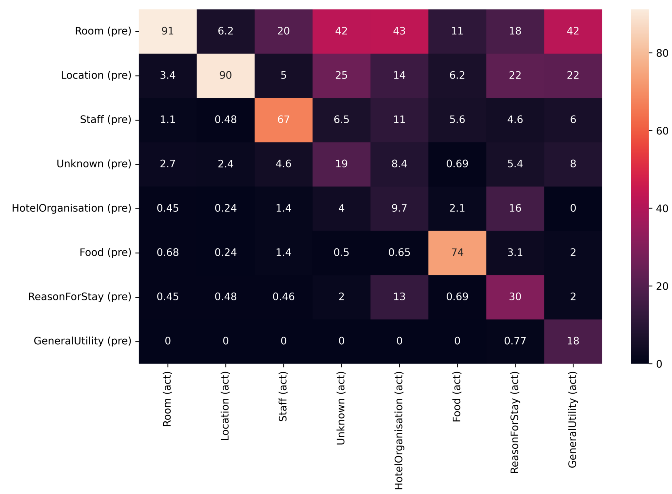


Distribution of the F1-Score

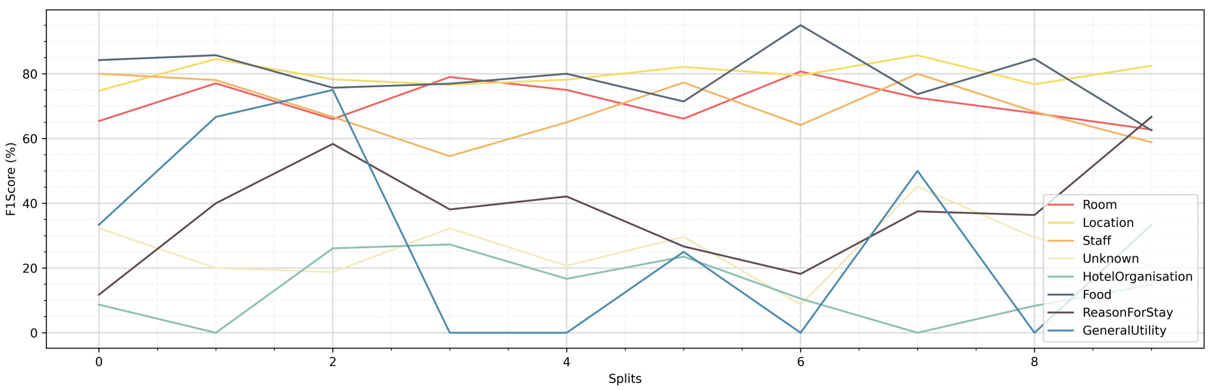
ConfusionMatrix:



Normalised ConfusionMatrix:



F1 Socre by split:



F1-Score per split