**Predicting restaurant review / sentiment classification / rating based on review text using translated and original, polish text.**

1. Introduction

Purpose of the projects

Given high quality of available tools, we investigate a particular strategy – parsing \_\_

The hypothesis is that machine translation will remove spelling and grammar mistakes, thus yielding higher quality training data.

Main assumptions / hypothesis

Rather than using polish reviews directly

1. Literature review

<https://link.springer.com/article/10.1007/s10462-020-09873-y>

1. Data

Data for

Finding the restaurants

Scraping the reviews

Filtering the reviews (language detection)

Preparing the data

Text normalization

* Tokenization
* Punctuation removal
* Stop-word removal
* Lemmatization

Feature engineering

* Tf-idf matrix

1. Modeling

Count of words

Multinomial Naïve Bayes Classifier

Binary Naïve Bayes Classifier, where the word frequencies are clipped – does not improve the results

Tf-Idf matrix doesn’t improve the result

Bernoulli

XGBoost

GRU

Why binary not tf-idf:

Also, very short texts are likely to have noisy tf–idf values while the binary occurrence info is more stable.

Given sentiment analysis task

In spite of their apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many real-world situations, famously document classification and spam filtering. They require a small amount of training data to estimate the necessary parameters. (For theoretical reasons why naive Bayes works well, and on which types of data it does, see the references below.)

Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution. This in turn helps to alleviate problems stemming from the curse of dimensionality.

* Word occurrence matters more for sentiment classification than word frequency

Embeddings

Recurrent neural network (GRU)

XGBoost

1. Results
2. Conclusions