Yelp review dataset

Branislav Doubek

March 7, 2020

Table of contents

1. Machine learning
2. Categorization
   - Sklearn
   - MIL
   - Keras
3. Classification
   - Sklearn
   - Keras
4. Classification - Keras performance
5. Results

Machine learning

- Text classification
  - Predict number of stars based on the review's text
  - E.g.: 'This restaurant was horrible' - negative review
  - Information about what customers feel is good/bad about the business
  - Positive review: user awarded 4 stars or more
  - Negative review: user awarded 2 stars or less
- Text categorization
  - Predict category of individual reviews based on review's text
  - E.g : 'Great customer service and cheap transport' - shopping review
  - Automatic tagging on documents
  - We train our models on shopping and restaurant categories

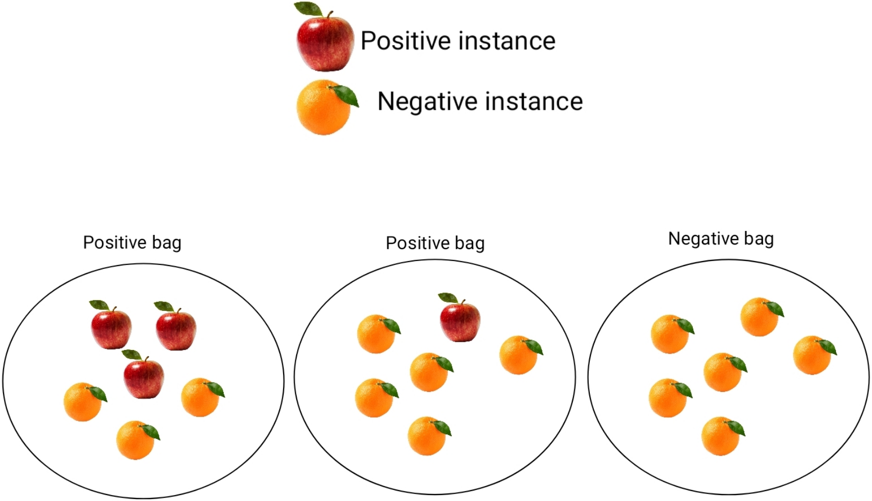
Categorization - Sklearn

- Each instance is calculated as mean of all word2vec representation of individual words inside a review
- Instance are then fitted to classical SVM classifier

Categorization - MIL

- If at least 1 instance in a bag is labeled as positive, whole bag is positive, otherwise it is negative
- Bag - Review
- Instance - Individual sentences
- Each instance is generated by calculating mean of vectors generated by word2vec from words inside a sentence
- We then use mi-SVM algorithm for classification on individual bags and their labels
- Based on work proposed in [1]

Categorization - MIL



Positive instance

Negative instance

Positive bag

Positive bag

Negative bag

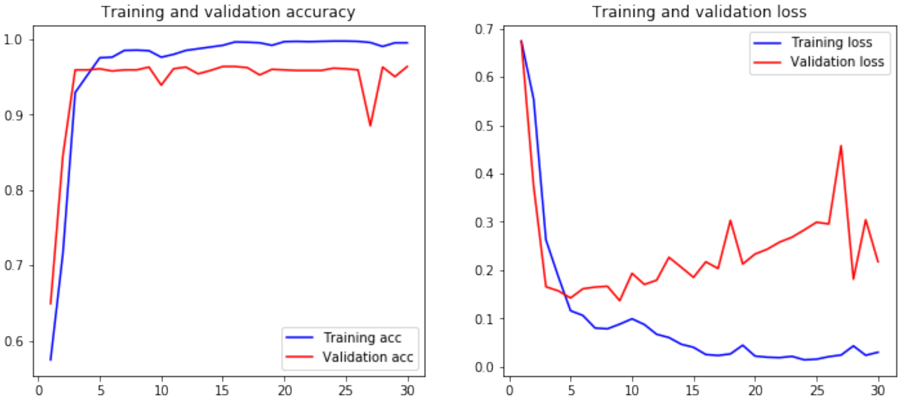
Categorization - Keras

- We implement same structure for classification as well as categorization
- We use precalculated vectors from glove implementation to transform individual words into vectors based on their semantic representation
- Embedding dimension 100
- Black box

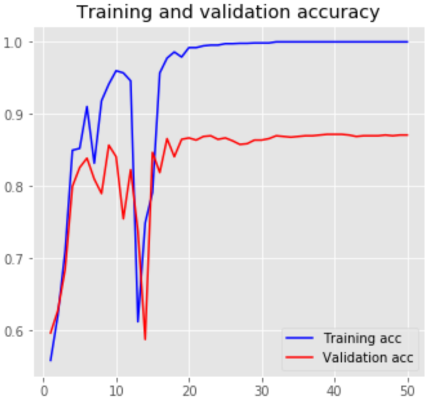
Classification - Sklearn

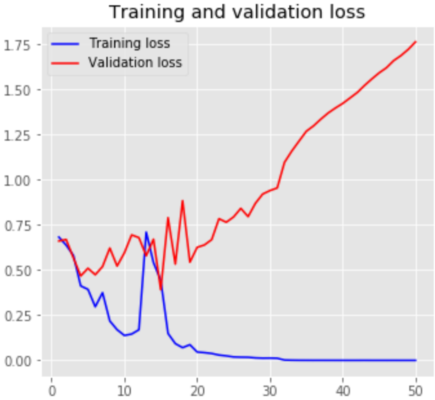
- Transform reviews with TF-IDF tokenizer
- Individual words are then passed to SVM

Keras loss for review categorization



Keras loss for review rating classification





Results

Review rating classification

- Train model to predict review rating from text
- Sklearn implementation - 0.876 accuracy
- Keras implementation - 0.871 accuracy

Review categorization

- Train model to predict review category fr om text
- Sklearn - 0.725 accuracy
- MIL - 0.6375 accuracy
- Keras - 0.9645 accuracy

References

[1] - Zhang, Multiple-instance learning for text categorization based on semantic representation (2017)