NLP Course Work

Egor Yurov

December 2022

Abstract

In this course work NLP text classification task was solved on movie reviews dataset. Project code: `https://github.com/Lab00001/test/`.

1 Introduction

Solving text classification problems is very important in modern life because it can be applied in many areas. For example in toxic comments classification which can help creating a friendly atmosphere on social networks or building films recommendation systems. In this work two types of vectorizers were compared using logistic regression classifier.

1.1 Team

Yurov Egor prepared this document.

2 Related Work

One of the first steps in solving text classification tasks is to create vector embedding of each document. There are many ways to do it: one-hot-encoding, bag of words [Harris, 1954], tf-idf vectorization.

After vectorization is finished this task may be solved using different classifiers. In this work logistic regression was chosen.

3 Model Description

Several types of vector representations were formed (bag of words, TF-IDF) for each representation a prediction matrix was created. After that for each matrix a logistic regression algorithm was launched and accuracy quality metrics was calculated.

4 Dataset

The dataset is called "movie_reviews". It is a collection of 2,000 movie reviews from IMDB that are labeled as a positive or negative review. This dataset can be downloaded using nltk.downloader. It contains 2000 reviews, each has a class label "0" or "1".

5 Experiments

5.1 Metrics

Metric used to evaluate this approach is accuracy (number of correctly predicted classes divided by total amount of objects in test dataset).

5.2 Experiment Setup

In this work two runs were conducted - one with BoW representation, another with TF-IDF representation. Each time logistic regression was launched with the following parameters: l2-regularization, binary mode. The dataset was divided into two parts: train and test as 2:1. Train part contained 1340 objects and test part contained 660 objects.

5.3 Baselines

As a baseline, TF-IDF with logistic regression was chosen.

6 Results

After creating vectorized representation each vector had dimention equal to 38864.

In BoW approach the accuracy value is equal to 99.9%. In TF-IDF approach the accuracy value is equal to 97.4%.

7 Conclusion

A dataset suitable for the task of text classification was found. Several word embeddings with logistic regression were tested. The algorithm with the BoW vector representation showed the best result compared to the basic solution.

References

[Harris, 1954] Harris, Z. (1954). Distributional structure. *Word*.