Staradubets Andrei

COMPARATIVE ANALYSIS OF MACHINE AND DEEP LEARNING ALGORITHMS IN SEMANTIC ANALYSIS

The article compare different approaches as machine learning and deep learning in semantic analysis of phone reviews in Russian.

Introduction

Nowadays, when you can get or lose millions on the market just in one second, it's extremely important to receive feedback for all your actions and products. But you can have no time to read every message about your new phone model. That is a time when sentiment analysis come to the spotlight and help you to see general report of (in our case) feedbacks.

As Indicator of correct work of algorithm has been chosen late submission for Kaggle competition [1] with accuracy score as a result. Since vocabulary of example reviews is connected with phones and provided in Russian, reviews from site 4pda.ru has been chosen as initial dataset for training of our models.

I. Data Quality Assessment

The main challenge of data cleaning in our case is that reviews are marked from 1 to 10 while Kaggle competition divides them only as positive or negative. The answer to the question "what mark is the highest negative?" is one of hyperparameters of the algorithm. After checking the balance of marks in the dataset (Fig. 1), exploring text of the reviews and testing with different values, it was found, that people usually write negative reviews with marks below 8. This value was chosen as a division into positive and negative reviews.

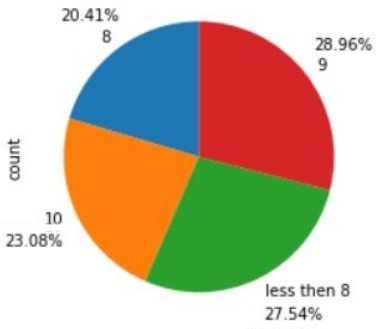

Рис. 1 – The balance of marks in dataset

The next challenge was to clean the reviews from punctuation and other special symbols. After cleaning the data the accuracy on train data of some algorithms has raised from 0.4 to 0.7.

II. Models and Algorithms

In sentiment analysis there is at least two different approaches: classic Machine Learning and Neural Network. We've tried both. As ML algorithm has been chosen implemented in python library sklearn tf-idf due to its best performance on our train data and on the Sentiment field in total, with LinearSVC for the same reason. After greed search has been chosen default parameters. Sklearn shows good performance and that is easy to build ML algorithms with that wide choice of already implemented functions in it.

As NN algorithm has been chosen LSTM Neural Network with 2 layers and size of hidden dimension equal to 512 implemented on PyTorch due to ability of it to remember context of current word. PyTorch is a powerful Deep Learning library for python that allows us to build networks in easy way, but still have an opportunity to tune a lot of parameters and structure of the model.

III. Feature Engineering

Since tf-idf requires only sequence of words as input feature, the only necessary thing with train and test dataset was cleaning data from extra symbols.

For LSTM NN due it was necessary to transform the reviews to arrays of encoded words with the same length of sentence. For that we transform reviews to array of words and after building vocabulary, encode each word with ID.

IV. Performance and conclusion

For Machine Learning Algorithm has been achieved 0.94 accuracy on test dataset, when for Deep Learning Algorithm – 0.64 accuracy.

1. Kaggle Competition [Electronic resource] – Mode of access : https://www.kaggle.com/c/product-reviews-sentiment-analysis/ – Date of access: 29.02.2020

Стародубец Андрей Сергеевич, студент 3-го курса ФИТиУ БГУИР, astarodubets@mail.ru

Научный руководитель: Трофимович Алексей Фёдорович, старший преподаватель кафедры ИТАС, trofimaf@bsuir.by