Machine Learning Text Classification Model with NLP Approach

Maria Razno[0000-0003-3356-5027]

National Technical University "Kharkiv Polytechnic Institute",
Pushkinska str., 79/2, Kharkiv, Ukraine

mari.razno@gmail.com

Abstract. This article describes the relevance of the word processing task that is written in human language by the methods of Machine Learning and NLP approach, that can be used on Python programming language. It also portrays the concept of Machine Learning, its main varieties and the most popular Pythonpackages and libraries for working with text data using Machine Learning methods. The concept of NLP and the most popular python packages are also presented in the article. The machine learning classification model algorithm based on the text processing is introduced in the article. It shows how to use classification machine learning and NLP methods in practice.

Keywords: Machine learning, Python, Pandas, Text classification, NLP, NLTK, Scikit-learn, Artificial Intelligence, Python Library, Deep Learning Texts

Over the last few years machine learning and artificial intelligence have become very hot topics. Nowadays their methods and approaches are a part of a huge amount of products, moreover it is a necessary thing in most applications and appliances. An example of using ML (Machine Learning) can be the automatic determination of important emails and quick responses in Gmail. Nowadays we can confidently say that and artificial intelligence with machine learning can push a person out of many technological processes.

Machine learning is the scientific study of algorithms and statistical methods that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model of sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. There are five types of machine learning algorithms: supervised, semi-supervised, active learning, reinforcement and unsupervised learning [1].

Natural language processing is a subfield of computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular, how to program computers in order to process and analyze large amounts of natural language data. Tasks in natural

COLINS'2019, Volume II: Workshop. Kharkiv, Ukraine, April 18-19, 2019, ISSN 2523-4013 http://colins.in.ua, online

language processing frequently involve speech recognition, natural language understanding, and natural language generation.

Text classification is one of the most important and typical task in supervised machine learning. Assigning categories of documents, which can be a web page, library book, media articles, gallery etc. has many applications like spam filtering, email routing, sentiment analysis etc. We would like to demonstrate how we can do text classification using the most common python machine learning and natural language processing packages like: Pandas, Scikit-learn, Numpy and little bit of NLTK.

In our study, we are creating the model, that will be able to classify user`s comment and give it a star rate from 1 to 5. Supervised machine learning requires to have prepared labeled data, so we use Yelp_academic_dataset_review in json format. We downloaded the dataset via the link: https://www.kaggle.com/yelp-dataset/yelp-dataset#yelp_academic _dataset_review.json. We got a lot of necessary tools by using Pandas library, that helped us to store data in convenient table, the columns of which are classification parameters and the rows – information for each classified object. This form of data storage is very effective in our study, especially for further accessing a particular column of data during the text processing [2].

The next step is to use natural language processing methods to normalize the text data. During our work we realized, that the package of libraries NLTK(Natural Language Toolkit) is great for our purpose. Thanks to its methods we removed all the stop words, that were not necessary for further analysis, from the text data. Also we needed to use text stemming in order to remove morphological affixes from the text. All of those step helped the model to make an accurate analysis of the text data and get the best clear features for the future classification [3].

The next step of our study was building the machine learning model. We used Scikit-learn due to the fact, that it is a wonderful library with a huge amount of opportunities. It has various types of analysis, moreover, it is the most convenient way of forming a model, because it provides a single interface for all conversion steps and the final result. Instead of using "Bag of words" approach and counting of each word in our text data, we use the tf-idf method for each pair of words in our reviews. Tf-idf normalizes the count by dividing the total sum of the meeting of a certain pair of words into the number of reviews in which these words appear. In such way, we get the model, that will find the most common words for each star rate, in other words it will get the appropriate features and select the best ones. As the result of our study, the model will be able to analyze user`s comment according to found features.

To summarize, in the course of our research, we can say that Python is a wonderful programming language, which provides a lot of great libraries for creating powerful machine learning models and proper natural language processing. For the task of building a machinelearning text classification model with NLP approach, we have reviewed the most popular machine learning libraries like : Pandas, Scikit-learn, Numpy, NLTK and built the text classification model with NLP approach. At the end of our study we get the learning model, that gives user answer with the appropriate star rate to user, according to his comment, and the list of the most common words for each star rate.

References

1. Langley, P.: Human and machine learning.Machine Learning,1, pp. 243–248 (1986)
2. Masch, C.: Text classification with Convolution Neural Net-works on Yelp, IMDB & sentence polarity dataset, https://github.com/cmasch/cnn-text-classification,24/02/2019.
3. Moschitti, A., Basili, R.: Complex Linguistic Features for Text Classification: A Comprehensive Study. In: Lecture Notes in Computer Science vol. 2997, pp. 181-196, Springer Science + Business Media (2004)