A Study of Text Classification using Python in AWS SageMaker

In every field of today's world, individuals must deal with very large amount of text documents. And the classification of documents/texts is an important task of supervised machine learning approach. Typically, most of the text data are profound in web, library book, media articles, bulletin board, printed news etc. Categorizing the documents from these sources can have applications like spam filtering, email routing, sentiment analysis etc. [1]. Also, the demand of topic modeling of the raw texts is also in demand with the exponential growth of text data resulting due to development of Internet and web-based applications.

In this project, we will use the 20 news group data available with in the scikit learn library [2]. And much of the task of the project will be performed in notebook instance of AWS SageMaker [3]. We will be using the module related to natural language processing like nltk [4] and other popular modules like NumPy and pandas.

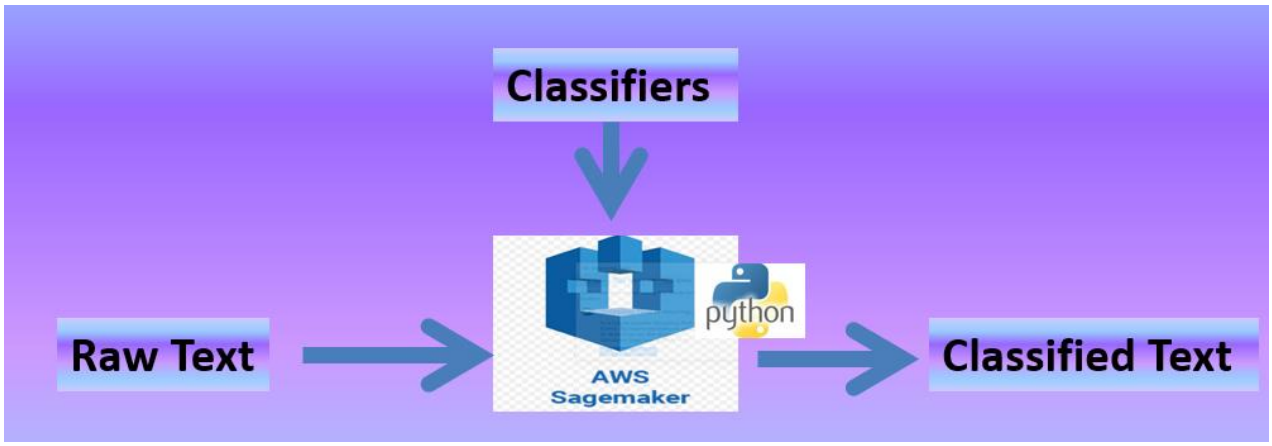

Figure 1. An Outline of Text Classification with Python in SageMaker

The outline of the main task of the project is as shown in Fig. 1. This project initiates by setting up the environment in AWS SageMaker that involves creating of a jupyter notebook instance. Next step is to load data set into the opened jupyter notebook. As text files are ordered series of words, we must convert these into numerical feature vectors before running machine

learning algorithms. With the help of scikit-learn we will be creating the feature vectors. The next move during our project is about running machine learning algorithms. Although, there are different algorithms available that can be used in text classification, for this project we have planned to use two popular algorithms, "Naïve Bayes" and Support Vector Machine [5]. Thereafter, we will apply the general machine learning approach which is splitting of the dataset into train and test, learning a model from the training data and then evaluating the learned model with the test data. During this study we plan to evaluate the learned models based on accuracy, precision, recall and score [6]. The reason behind using these metrics is that we are perform a classification task and these are popular metrics for evaluating classification models.

With the time permitting for the project, we will try to use a inbuilt AWS SageMaker model to perform the topic modeling of the text data. We will also introduce the concept of pipeline with the help of scikit learn. Also, the project is initiated with the plan to involve some interesting data analysis other than text classification. The primary expected result from this project is to dig into the capability of natural language processing algorithm used and to have a clean report on how the whole process of text classification works.

References

[1] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text Classification Using Machine Learning Techniques."

[2] F. Pedregosa FABIANPEDREGOSA *et al.*, "Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011, Accessed: Jan. 12, 2022. [Online]. Available: http://scikit-learn.sourceforge.net.

[3] Sagemaker, "Amazon SageMaker - Machine Learning Solution." https://aws.amazon.com/pm/sagemaker/?trk=ps_a134p000007BxaEAAS&trkCampaign=acq_paid_search_brand&sc_channel=PS&sc_campaign=acquisition_US&sc_publisher=Google&sc_category=Machine Learning&sc_country=US&sc_geo=NAMER&sc_outcome=acq&sc_detail=%2Bsagemaker&sc_ (accessed Jan. 12, 2022).

[4] M. M.L, "A Gentle Introduction to the Bag-of-Words Model." https://machinelearningmastery.com/gentle-introduction-bag-words-model/ (accessed Jan. 12,

2022).

[5] V. Jakkula, “Tutorial on Support Vector Machine (SVM).”

[6] G. Schröder, M. Thiele, and W. Lehner, “Setting Goals and Choosing Metrics for Recommender System Evaluations,” Accessed: Nov. 19, 2021. [Online]. Available: http://www.grouplens.org/node/73.