Performance Evaluation on Vector-Model Combinations

Project Proposal B

GitHub link: https://github.com/trela47/PEoVMC_nlp-project

Team Members:

Yamuna Bollepalli

Saisri Teja Pepeti

Saikiran Yedulla

Blessy Kuriakose

Goals and Objectives

Motivation:

Exploring the effectiveness of differing vectorization-and-model combos in the text based prediction task of spam detection. By this project we will be able to accomplish which vectorization will work with best model accuracy and least time complexity by comparing combinations.

Significance:

Initially we will be collecting a data set for example email spam detection. We will be cleaning the data set with basic nlp techniques such as stemming and lemmatization and then we apply different methods to accomplish vectorization such as

1. Bag of words
2. L1 Normalized term Frequency
3. L2 Normalized TF-IDF
4. Word2vec

Then we supply different machine learning classification models such as

1. Naive Bayes
2. Random Forest classification
3. Support Vector Machine classification
4. Decision Tree classification

Then we will conclude the best vectorization method and machine learning model combo. By considering metrics like accuracy, confusion matrix or classification report and time complexity.

Objectives:

- Clean the raw text.
- Get language-based vectors from the cleaned text.
- Make the models for training and testing and track the metrics (confusion matrix, time complexity, etc.)
- Try each vector for each model.
- Evaluate the metrics to conclude the best combination.
- Calculate the time complexity.

Features:

We will be using the spam detection dataset from kaggle, of which the direct features are available as direct text. Alterations of cleaning and vectorization (frequency-based understanding of the words present) will be completed prior to use with the models.Then evaluate the model performance based on time complexity and model metrics.

Project lifecycle:

Figure 1: Project Lifecycle.

References

https://medium.com/analytics-vidhya/magic-of-tf-idf-202649d39c2f

https://neptune.ai/blog/vectorization-techniques-in-nlp-guide

https://www.sketchbubble.com/en/presentation-project-life-cycle.html

https://machine-learning.paperspace.com/wiki/machine-learning-models-explained