Cloud Computing for Data Analysis Fall 2022 Project Proposal (Revised)

Title: ***Sentiment Analysis of Social Media Responses (Twitter) With Thorough Explanation Using Spark and Python***

Rezaur Rashid

Introduction

With the popularity of Big-data in recent years, we are experiencing rapid growth in data collection and they are becoming more available publicly and their usage in various applications is omnipresent which is creating new challenges in researcher communities [1]. Although several machine learning tools are being developed to deal real-time data processing, most of the traditional tools processes data in batches and struggle handling large-scale data.

Therefore, a cloud computing tools such as Spark, can provide the option to process such large-scale data as well as real-time computation in less time. Such tools can help large-scale data preprocessing, cleaning and aggregational computations.

Project Overview

For this project, we will implement a multiclass classification version of the Support Vector Machine (SVM) model to perform sentiment analysis of the user responses in social media (twitter) which has multiple classes. We know Apache Spark MLlib has two linear classification model: SVM and Logistic Regression (LR) and although LR can do multiclass classification, the MLlib SVM can only do binary classification. Therefore, the goal of this project is to develop a multiclass SVM classifier in a distributed fashion and also compare the model with LR classifier model.

Dataset

For our project, we will use the "Sentiment140 dataset with 1.6 million tweets" dataset from Kaggle [2]. This dataset provides user responses to different products, brands, or topics through user tweets on the social media platform-Twitter. The dataset was collected using the Twitter API and contained around 1,60,000 tweets. This dataset contains 6 attributes where two of them are polarity score (categorical) and text (tweets) we are mostly interested in.

Project Motivation

In a world of social-media and businesses, where people are expressing their emotional states, sentiment analysis plays a big role. It is an NLP application that helps to understand human behavior through text and speech. For example, using sentiment analysis tools, business organizations can learn about the emotional and behavioral state of customer using machine learning models which can enhance customer relationship and service.

And in the age of continuous data generation, cloud computing tools like Spark can help processing this large-scale real-world data with simple API, fast and distributed processing.

Proposed Work

Our proposed work is divided into several steps.

1. Data cleaning and data preprocessing in AWS cluster
2. Feature embedding/transformation of Textual features using MLlib such as Word2vec
3. Implement the multiclass classification SVM model: train, validate, test and model
4. Compare results with the Logistic Regression Model
5. Get feature importance for explainability
6. Leveraging the work of [3], developing a similar criterion to estimate the causal-effects of the features

Deliverables

I. A sentiment analysis multiclass classifier using Spark and Python and their thorough explanation. At the end of the project we will have a multiclass SVM classifier developed in the distributed format. Some performance analysis of the implemented model extending to comparing with Logistic Regression multiclass classification from Apache Spark MLlib. **(must accomplish)**

II. Thorugh Explanation.
   a. Explaining the model behavior in terms of important features that drive the model predictions.
   b. Causal Feature Selection. Finding the causal factors (features) that has high impact on the target/prediction using the average causal-effect estimation method. **(like to accomplish)**

III. Sentiment analysis of real-time stream data using Spark Structured Streaming tools. Also, extracting different graph centric features such as from the user and post interaction we can create a user network and then extract graph properties like number of replies, reposts, parents of users' polarity, children of users' polarity etc. **(would ideally like to accomplish)**

References

1. Zhai, Y., Ong, Y. S., & Tsang, I. W. (2014). The emerging" big dimensionality". IEEE Computational Intelligence Magazine, 9(3), 14-26.
2. Dataset https://www.kaggle.com/datasets/kazanova/sentiment140
3. Panda, P., Kancheti, S. S., & Balasubramanian, V. N. (2021). Instance-wise Causal Feature Selection for Model Interpretation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1756-1759).