Cloud Computing for Data Analysis Fall 2022 Project Proposal

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.

For this project, we will implement a variant of recurrent neural networks (RNN) model using Spark and Python to perform sentiment analysis of the user responses in social media (twitter) and provide a thorough explanation (e.g. causal factor, important features) behind the sentiments of users.

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.

Motivation

In a world of social-media and businesses, where people are expressing their emotional states, sentiment analysis plays a big role. It is a 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. Download the data and upload it to AWS cluster
2. Data cleaning and data preprocessing for modeling
3. Apply LSTM networks model for sentiment analysis

4. Apply SHAP to get explanation for important features in sentiment prediction
5. Apply causal model (e.g. LINGAM) to get causal explanation

Deliverables

a) A sentiment analysis model using Spark and Python and their thorough explanation (must accomplish)
b) Feature synthesis. Extracting different set of features (e.g. graph centric properties) from user interactions instead of just word/vector embeddings from text for the model (like to accomplish)
c) Sentiment analysis of real-time stream data using Spark Structured Streaming tools (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