Individual Report

Prabas Leti UOH

December 28, 2023

Contents

- Preprocessing (Thunderbird log dataset)
  Preprocessing is Sushanth's part i have helped him with the thunderbird dataset which has further helped him for spirit dataset where both use fixed-window grouping.
- Bert base Model
  I have helped Mahendar in building the bert base model.
- Training of ThunderBird Dataset.

Thunderbird dataset

- Thunderbird datatset : The log data contains normal and abnormal messages which are manually identified. Thunderbird is a large dataset of more than 200 million log messages. We leverage 10 million continuous log lines for computation-time purposes, which contain 4,934 abnormal log messages 0.49
- Finding Data
  This is the link : https://zenodo.org/records/8115559.
- The above provided link gives the Parse data. Log Hub only provided 2k lines of raw data, which is not adequate, hence I've used 100k or more entries of log data available at LogPai. Raw data is parsed using logparser python library and we use, Drain parser is primarily used.
- Datasets represent highly imbalanced class distributions with anomaly ratios that can be only 0.1

Log Grouping

- I have used fixed-window grouping for this dataset. Session window-grouping is not used for this dataset because it does not have blocks or nodes.
- Fixed Window, Timestamp, window size: 900sec(15min).
- First Converted the 'Timestamp' column to datetime format.
- Defined the interval size in seconds (1800 seconds=30 minutes).
- Created a new column for the window index.
- Grouped the data based on the 'WindowIndex'.
- Initialized a list to store results,Created a new DataFrame from the list of dictionaries.

Labelling Sequence

- A log sequence is abnormal if it contains an anomalous log message.
- The general idea: "If all the events in a sequence are Normal, then the sequence is labelled 'Normal' else 'Anomalous' "
- This sequence file is generated for the dataset and sent to Mahendar for training BERT model.

Bert Base Model

- BERT, which stands for Bidirectional Encoder Representations from Transformers is a transformer-based model.The base model refers to the smaller version of BERT, as compared to larger variants like BERT Large.
- But this model was built in the beginning using Tenserflow datasets, the bert model and tenserflow were not compatible. Which eventually got State Management Problem. Both in Bert base model and large model.
- Later mahendar has created another BertforSequenceClassification model where he used pytorch datasets. Which did not have any state management problems. Training is done correctly and also the classification is done at this level. Which gave the final output Anomalous or Normal.

Training

- Repeated the training of Thunderbird Dataset as mahendar did for other datasets.