1. Problem Statement -- Riya
2. Project Architecture -- Riya /Vijay
3. HLD – Training Pipeline , Prediction Pipeline - Riya /Vijay
4. LLD -Subcomponents flowchart as below
5. Tools to be used – CICD , VScode,MongoDB ,
6. EDA --Riya/Vijya
7. Project Component structuring. --Riya
8. Code practices
9. Data Ingestion - Riya
10. Data Validation - Vijay
11. Data Transformation -Riya/Vijay
12. Model Training - Riya/Vijay
13. Model Eval -Riya
14. Model Pusher -Vijay
15. Prediction Pipeline -Riya/Vijay

Mongo DB install – Vijay homework

Code push API, Git for only data pipeline, Code modularization – Riya homework.

Minutes of meeting -> 4th Feb 2023

1.Data collection - > kafka read data and dump json in mongo db - repository github ->big data team

2.Training repo ->

data ingestion- create connection with mongo db and pull the latest records

valid -

eval -

create model

pusher

output - ECS cluster pickle

3.Prediction repo -> ML OPs -> production repo

pickle - run model - accuracy

json api -> parquet -> data ingestion -> pyspark df - > data validation ->transform ->model training (mongo db) ->model evaluation(aws model or new model with test file )? -> pickle file store it on AWS ->CICD (github actions)

MLOps CICD

CircleCI

airflow - >JD

sagemaker

1.Complete EDA for NEw york for Topic Modeling -> 7th Feb

hyperparameter ,if any other alogs for topic modeling,bert and transformer

2.Complete EDA for NEw york for Sentiment Analysis -> 8th Feb

CNN,LSTM and GRU, Bi directional LSTM

3.Project Architecture

4.cleanup json -> 8th feb connect on 8th feb

5.data ingestion and validation -> vijay and validation - 9 to 14th feb (checking the schema ,data preprocessing- remove cases ,punctutations ,)

6.data transformation -> lemmatization and tokenization, tf-idf , new features? - 13th feb

7.model training 14th feb connect and decide steps further ->

8.model evaluation 21st feb onwards - > compare production model versus freshly trained model

9.model push 28th feb

10.push to cloud - 4th /5th march

11. flowchart - 6th march

prediction pipeline(calls the pickle file of model) (test file -> ) -> crime/politics output

calls model and gives output for test file

* only calls production model and gives results for test file

Project architecture - flowchart

Flowchart for each data stage -