10 Academy

Week 0 - Report 2

Name: Yishak Tadele

Data Science Project Methodology

CRISP-DM

1. Business Understanding

    - The goal of this project is to analyse the Twitter data and generate sentimental analysis or topic modelling results.
    - Sentiment analysis is a deep learning technique based on natural language processing that identifies whether a statement is positive, negative or neutral. The sentiment analysis result can be used to gain a deeper insight into people's thoughts regarding the two huge nations (China and USA). Every tweet is analysed based on a variety of features(from metadata or the tweet content itself). The sentiment analysis can be used to know the general perspective of the society on the issues, and also allows the organization to categorize the tweets and target advertisements or beneficial offers and customize based on their respective response.
    - The second main target of this project is topic modelling. Topic modelling uses to identify the main topics that are trending and popularly spoken about. It can be helpful to categorize the main topics that are being said. And also knowing the popular topics can lead any organization to customize their branding and improve their design strategy according to the need of the majority of society.
    - The model to be developed in this project should be able to model the main topics that are being talked about in the dataset and analyse the sentiments of most of the users on those particular topics.
    - The third aim can be to study the relationship of the main topics with the geographical location feature. This can be helpful to understand how the topics vary according to geographical location changes.

2. Data Understanding

    - The dataset to be used in this project is two types

        a. Global dataset: this dataset is collected using keywords: *['chinaus', 'china Taiwan', 'chinaTaiwancrisis', 'Taiwan', 'XiJinping', 'US-CHINA', 'Pelosi', 'TaiwanStraitsCrisis', 'WWIII', 'pelosivisittotaiwan'].* The dataset is a JSON dump collected from Twitter developer API. It consists of metadata(author info, tweet info, retweet count, tweet contents and etc..) of each tweet that is made during the days August 6-7, 2022.

        b. Africa-based dataset: this dataset is the same as the above dataset but is collected with the additional feature of the location. This adds an extra feature to the dataset collected so each tweet author's original location is provided along.

3. Data Preparation

    - The dataset will be cleaned using cleaning modules that perform the following tasks

- Data Imputing: to handle missing and null values
- Outlier rejection: by performing mathematical analysis outlier samples are rejected or removed so the model will not be biased when trained.
- Standardization: makes the feature values to a common scale without destroying their relevant information.
- Dimensionality reduction: is the process of reducing the feature size based on its relevance to the model development. It is done because having a large feature set can be computationally expensive and could make the model diverge or result in the curse of dimensionality problems.

4. Modelling

- The problem can be modelled as a classification problem by using the sensitivity feature as a label. In this way, I can train a model that performs better at identifying a sentiment of a given tweet. The other approach is using clustering methods to identify the significant topics in the dataset. To detect data drift I can use a regression model and predict the geographical location according to the main topic found in a given tweet.
- This can be decided after exploring the dataset with a deeper understanding.

5. Evaluation

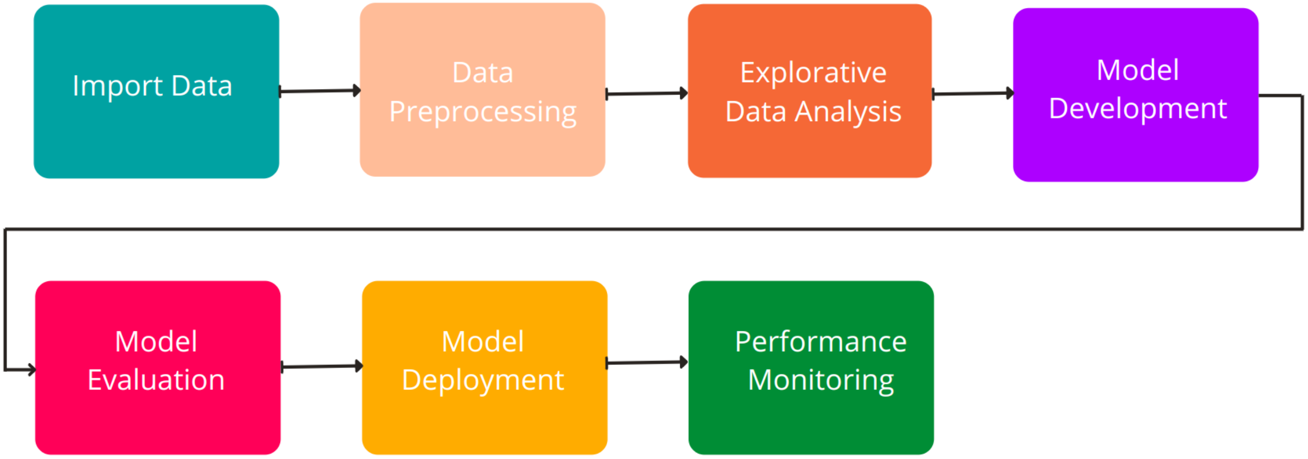
- In this stage, I will implement a variety of evaluation techniques to measure the performance of the developed model. If the problem can be modelled as a classification problem then accuracy, F1-Score, AUC and RMSE can be used to evaluate its performance.

6. Deployment

- The findings will be presented on the local server using a streamlit dashboard and documented report that will be submitted on the final day of this evaluation. A variety of plots will be included in the report along with my personal findings.

ML Workflow Chart and Description

Diagram



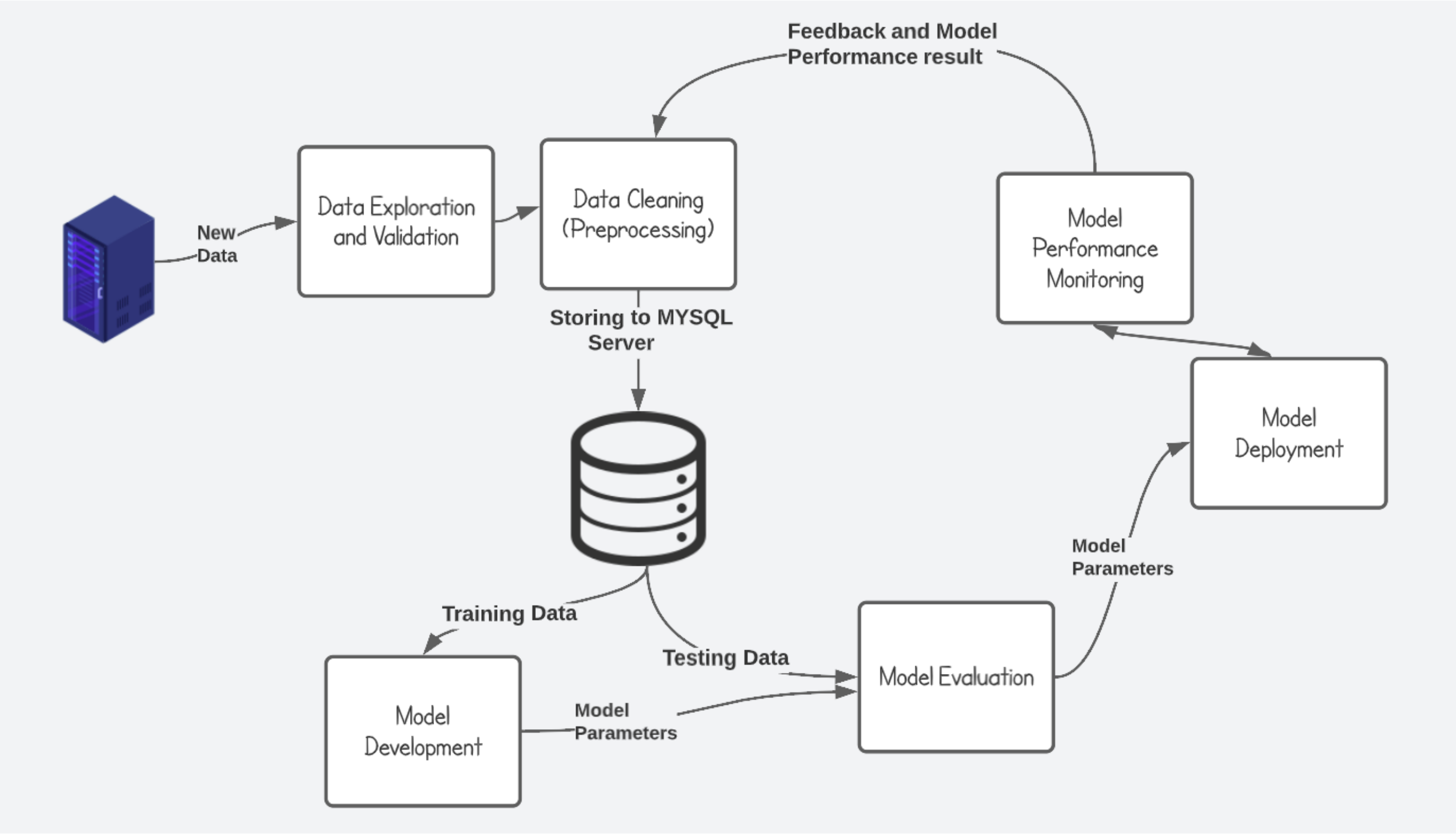
Description

1. Import Data
    - Read JSON file to Pandas Dataframe
        - Select the best attributes
    - Implement Data Consistency Check
    - Store in .pkl format for the next data pipeline
    - Populate MySQL
2. Data Preprocessing
    - Imputation, Normalization and Feature transformation and standardisation
    - Word cleaning
        - Lemmatization
        - Stop words
3. Explorative Data Analysis
    - Visualization
        - Matplotlib and Seaborn
        - Word clouds
    - Mathematical analysis and insight
        - mean, max, min, outliers, percentiles etc
        - Box plot, histograms, bar charts

   - Dimensionality reduction and feature engineering
      - Principal component analysis

4. Model Development
   - Model Training using
      - Scikit learn, Gemsi
      - K-means clustering
   - Topic modelling and sentiment analysis
5. Model Evaluation
   - Train-test split with cross-validation
   - Evaluation metrics: Accuracy, Precision, F1-Score, RMSE, AUC
6. Model Deployment
   - Develop an Integration code base
   - Load from MySQL
   - Present in Streamlit dashboard
7. Performance Monitoring
   - Deploy MLWatcher
   - Implement Data Drift Trigger models

MLOps Pipeline


New Data
Data Exploration and Validation
Data Cleaning (Preprocessing)
Feedback and Model Performance result
Model Performance Monitoring
Storing to MYSQL Server
Model Deployment
Training Data
Testing Data
Model Parameters
Model Development
Model Parameters
Model Evaluation