**Report on: Task 2 - Data Science Workflow & Engineering Pipeline**

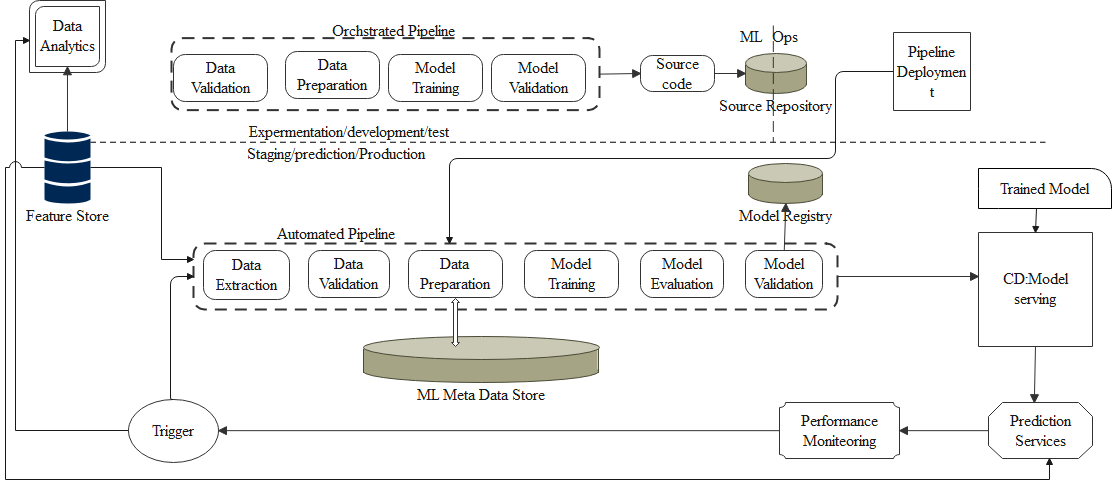
**Data Preprocessing:**

I have started by importing all the packages used to preprocess my data, including pandas, json, nltk, seaborn, matplotlib. Preprocessing has been installed in my Jupiter notebook and imported accordingly. Then, after I have written a code that loads my json file into my notebook and displays the content of the json file in the form of a data frame, I have seen detailed information about the data, including the data type, number of rows, number of columns, and shape of the data. Then I selected to view the first 5 files using df.head(5) to view the top 5 records from the json file. I removed all the special characters, emojis, and symbols from the file using regular expression functions. Finally, I have compared both the preprocessed and original data side by side by selecting the first three records. Finally, the preprocessed file is saved into a.csv file and loaded into the PostgreSQL database.

**Loading Data to Database:**

I used a PostgreSQL database to save my data from Panda's library to my local machine. I have installed and loaded psycopg2 for the database connectivity. Then I created a database named twitter-analysis on my PostgreSQL database and I provided all the privileges to access my dataset for writing and reading data from it.

**Workflow Diagram**

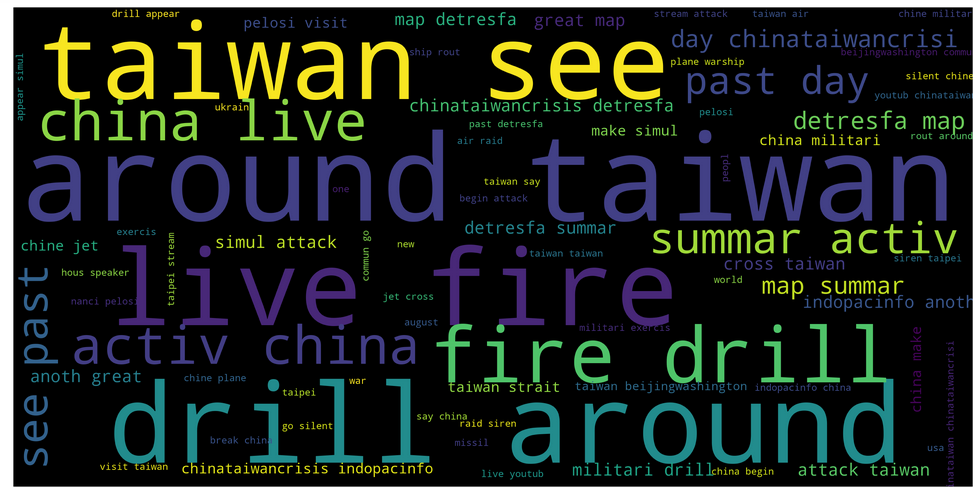


**Topic Modeling:**

I have used two main libriaries for this task namily gensim and pyLDAvis. I have used also word cloud library to print out the relationships and the most expressed word typed in the document.

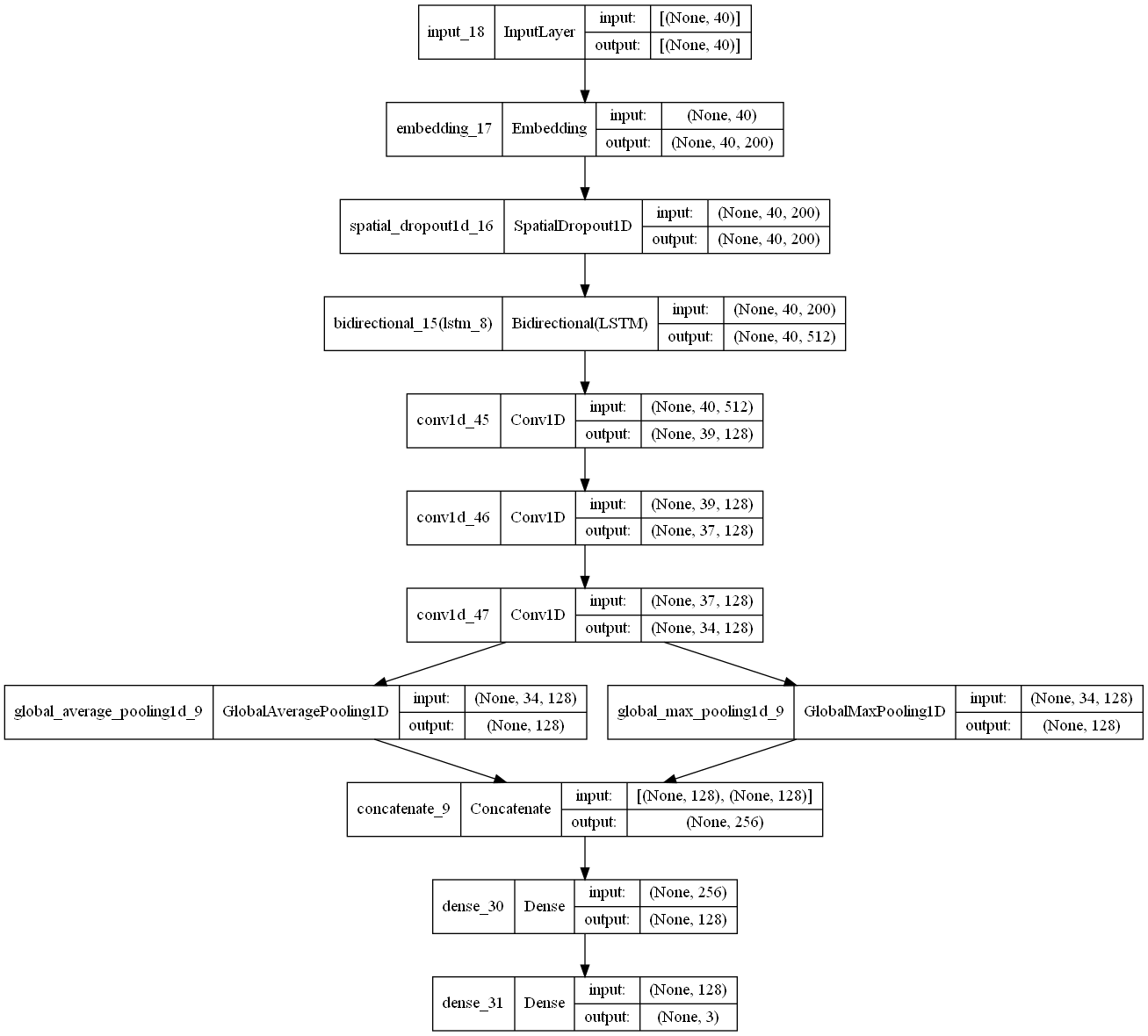
**Sentiment Analysis**:

I have preproces the data as usuall including finding the root word from the document listed and the polarity of the sentiment is categorized into three categories postive,nuteral and negative sentiments and the most named name is displayed in the word cloud as shown in the figure below.

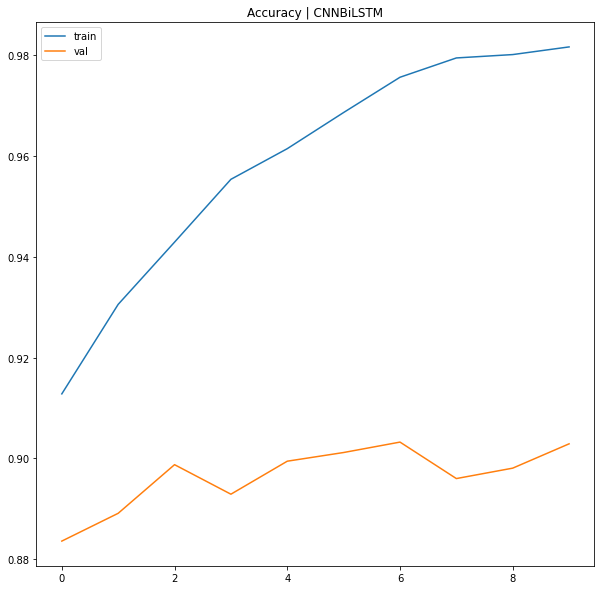


**Model Accuracy:**

Model Architecture –



Accuracy –



Confusion matrix and accuracy report:

