**TEAM NAME: Republic of Wadia (Kushaj Mallick, Nishant Rana, Swastik Mukherjee)**

*Placed 2nd at COMSYS Hackathon*

ARTICLE FOR TEXT CLASSIFICATION:

The BERT (Bidirectional Encoder Representations from Transformers) model was used to train and predict the mental illnesses based on the sentiments provided. The dataset was preprocessed by encoding the five available mental illnesses – Anger/IED, Anxiety Disorder, Depression, Narcissistic Disorder and Panic Disorder – by a respective integer and the dataset then was converted into a training dataset and a cross validation dataset with a split size of 0.2. The sentiments provided in the “Text” column were tokenized, converted to their respective TensorFlow format and finally input\_ids, attention\_masks and token\_type\_ids were added to this new TensorFlow data to be received by the BERT Classifier. We now declare a class called BERTForClassification that contains an \_\_init\_\_ method where we define the BERT model and a Dense layer with softmax activation along with a call method that takes the input data, extracts the pooled representation of the input sequence and passes this pooled representation through the Dense layer before returning its output. Finally a BERT Classifier is defined that passes the model to the class. It has a learning rate of 10^-5 and is used with the adam optimizer. Now that the model is trained, we predict on the cross validation dataset and the model is able to successfully perform the task with a 96.47% accuracy after 20 epochs of training. The actual testing dataset is taken and tokenized before passing to the trained model, which then outputs different weights of the encoded illnesses for each of the given sentiments. The highest value among the weights for each sentiment is taken to be the final result and other columns are discarded. The columns are decoded to their respective illness and added to their respective IDs.

ARTICLE FOR TRANSFER VALUE PREDICTION:

The given dataset contains 25 features and 1293 rows. After successful feature testing using Pearson Correlation where we test each of the features with the target value (Value at beginning of 2023/24 season), the two features having the lowest correlation with the target value (Name, Country) are discarded. The missing values in the data were filled by interpolation. The dataset is split into a training and a cross validation dataset with a split size of 0.2. We then apply a Linear Regression model on the training dataset and the model predicts on the cross validation dataset with an RMSE value of 2.03. The model predicts on the testing dataset after the redundant Name column is removed from it. The output NumPy array is then converted into its respective Pandas dataframe which is modified by adding the respective IDs.

ARTICLE FOR TIME SERIES ANALYSIS:

The SARIMAX (Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors) model was used for this problem statement. The data was preprocessed by removing the outliers and replacing them with their interpolated values. Two columns with variance threshold of 0 were removed and the Year and Month columns are combined into one column in a datetime format before being sorted in ascending order. Stationarity of the data through the Augmented Dickey-Fuller test is achieved by using a first value difference of CO2 levels – which tells us that the parameter “d” of the SARIMAX model will be 1. Next the Partial Autocorrelation (PACF) and Autocorrelation (ACF) graphs were plotted to determine the “p” and “q” parameters of the model, respectively but to avoid any confusion regarding optimal parameter choice we also compare the AIC and BIC values for the 2 sets of suitable parameters obtained from the graphs ((1,1,1) and (1,1,2)). The values thus obtained are extremely close to each other and thus the final result will be the weighted average of the models trained using these 2 sets of parameters. The dataset is divided into a training and cross validation dataset with a split size of 0.2 and both the SARIMAX models are trained on the training dataset. The models perform on the cross validation dataset with an RMSE value of 0.3596161 and 0.3593560, respectively. The models then predict values for the 24 months in the testing dataset and each respective value is combined by a 50/50 weightage from the two models to produce the final output.