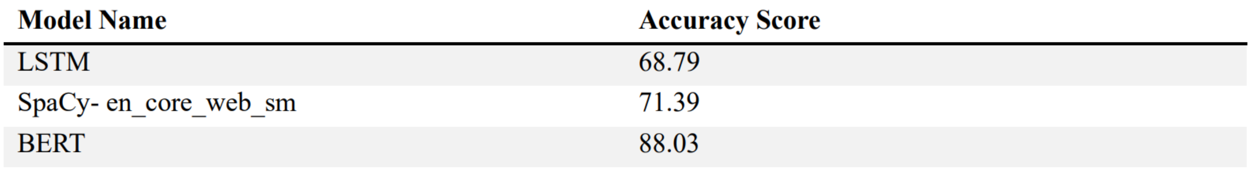
Jadavpur University, Kolkata, India and iLEAD, Kolkata (Dipan Mondal, Sabarna Saha and Debkumar Bera)

The submitted text classification method is the end result of comparison of two separate NLP models and choosing the best one. The first is a simple LSTM based model which needs some preprocessing. The data preprocessing phase includes Unwanted text removing, Decode contractions, Punctuation removing, Stopword removing and One-hot encoding(in alphabetical order) of the words in each sentences. The LSTM model which is a sequential one includes an embedding layer, LSTM layer, dense layers, and dropout layers. There is only one LSTM which has 300 neurons followed by a dense layer with 20 neurons, kernel_regularizer L2(1e-6) and ReLU activation with dropout of 0.3 followed by another dense layer with 20 neurons, kernel_regularizer L2(1e-6) and ReLU activation. Output layer has 5 neurons and softmax activation function. The model is the result of multiple trial and error, and some research from internet . It is then trained with ADAM optimizer with learning rate 0.02, considering the categorical crossentropy loss.

The second model is a Transformer based model with attention mask. Here BERT autotokenizer is used with trainable input mask. Before that preprocessing is done involving unwanted text removing, contractions decoding. The model involves an embedding layer based on BERT autotokenizer then an attention mask followed by a global maxpool layer. Followed by a dense layer with 128 neurons and ReLU activation with 0.1 dropout with another dense layer with 32 neurons and ReLU as activation function. The final output layer has 5 neurons with softmax activation function. Finally the BERT parameters are set as trainable for fine tuning. This model is trained with ADAM optimizer with learning rate 5e-5 and epsilon 1e-8 considering the categorical crossentropy.

A SpaCy-'en_core_web_sm' model is used as benchmark comparison.

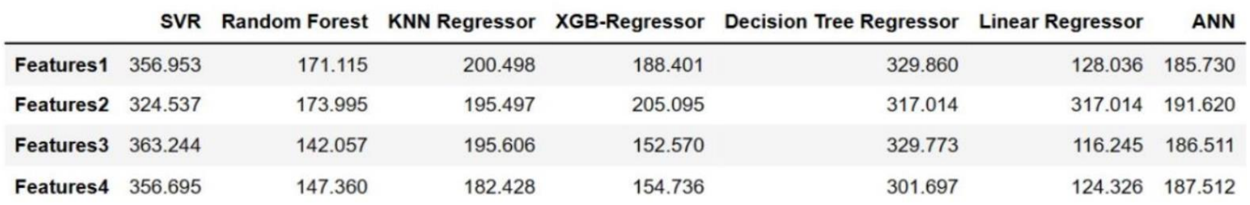


Model Name	Accuracy Score
LSTM	68.79
SpaCy- en_core_web_sm	71.39
BERT	88.03

In the Market Value Prediction problem a feature selection and multiple model test is used, feature selection is done using mutual information. Thus four batches of selected features are taken and multiple regression models are applied on each of the batches. First batch based on co-relation with co-relations > 0.3, second batch with co-relation greater than equal to 0.4, third batch with mutual information greater than equal to 0.05 and fourth batch with mutual information greater than 0.06.Batch i is represented as features i.

Since the standard deviation curves are majorly non-gaussian, standard scaler is used for feature scaling.

RandomForestRegressor, DecisionTreeRegressor, SVR, KNeighborsRegressor, and Artificial Neural Networks (ANN) models are used on each of the batches. Comparison is made based on Mean Squared Error.



	SVR	Random Forest	KNN Regressor	XGB-Regressor	Decision Tree Regressor	Linear Regressor	ANN
Features1	356.953	171.115	200.498	188.401	329.860	128.036	185.730
Features2	324.537	173.995	195.497	205.095	317.014	317.014	191.620
Features3	363.244	142.057	195.606	152.570	329.773	116.245	186.511
Features4	356.695	147.360	182.428	154.736	301.697	124.326	187.512

The carbon dioxide level prediction which is basically a time series data is predicted using two models and we have taken the best one. Before that the data preprocessing steps mainly includes outlier removing, the outliers and null values are replaced with the average immediate previous and following values. And the values are scaled using MinMax scaler for model optimization and simplicity in time series. The first model is a LSTM model with 256 neurons and a window size 30 with activation ReLU. Following a dense model with 1 neuron acting as a output layer, it is trained using ADAM optimizer with MSE loss function. The second one is the ARIMA model by analyzing ACF and PACG plots we have used an ARIMA model of order (9,2,9)

Model	MSE (Mean Squared Error)
ARIMA	1.357
LSTM	11.239