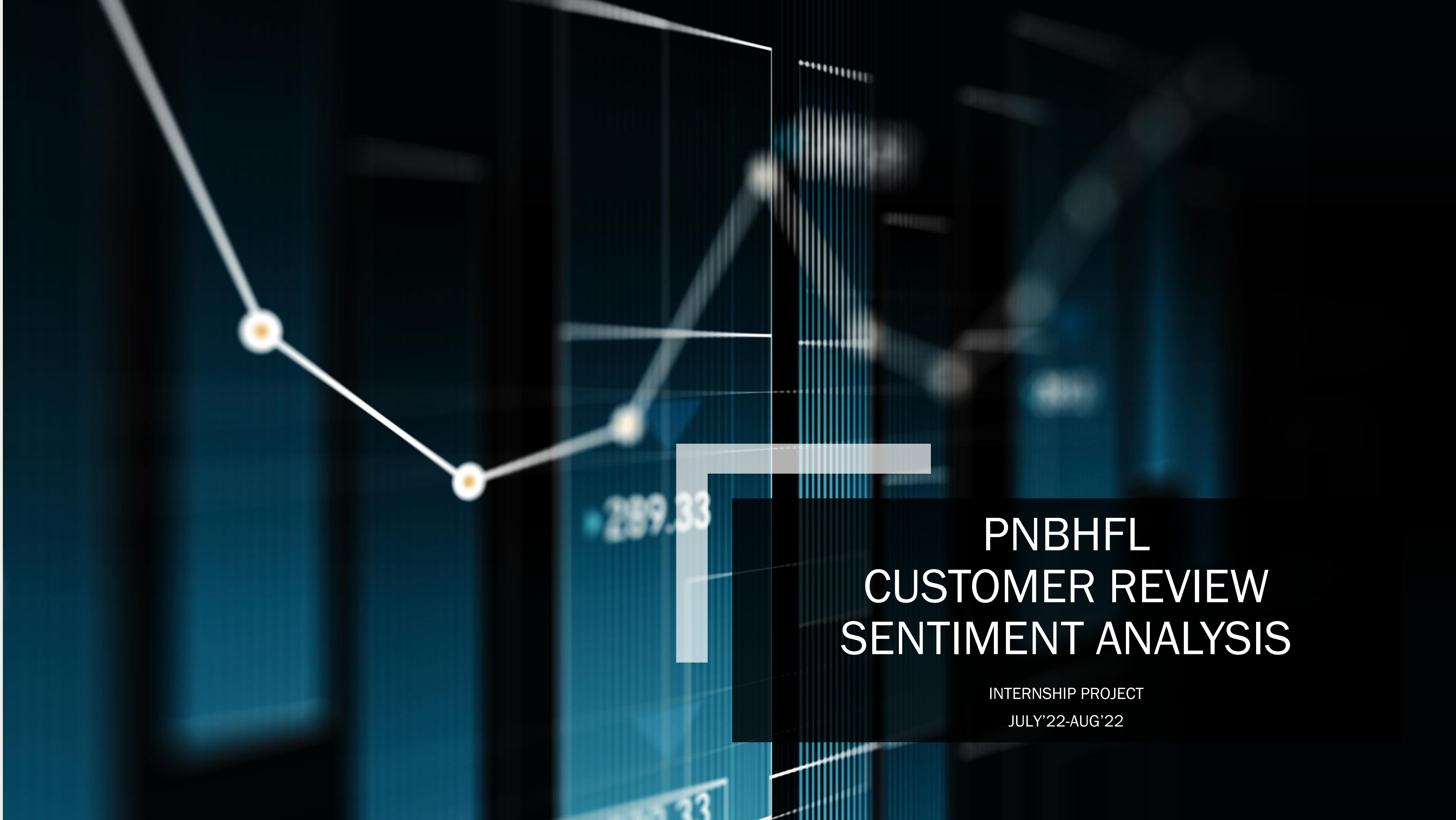


PNBHFL
CUSTOMER REVIEW
SENTIMENT ANALYSIS

INTERNSHIP PROJECT

JULY’22-AUG’22

Scope of This Customer Review Sentiment Analysis Project For PNBHFL

pnb Housing
Finance Limited
Ghar Ki Baat

ORGANIZATION 💼

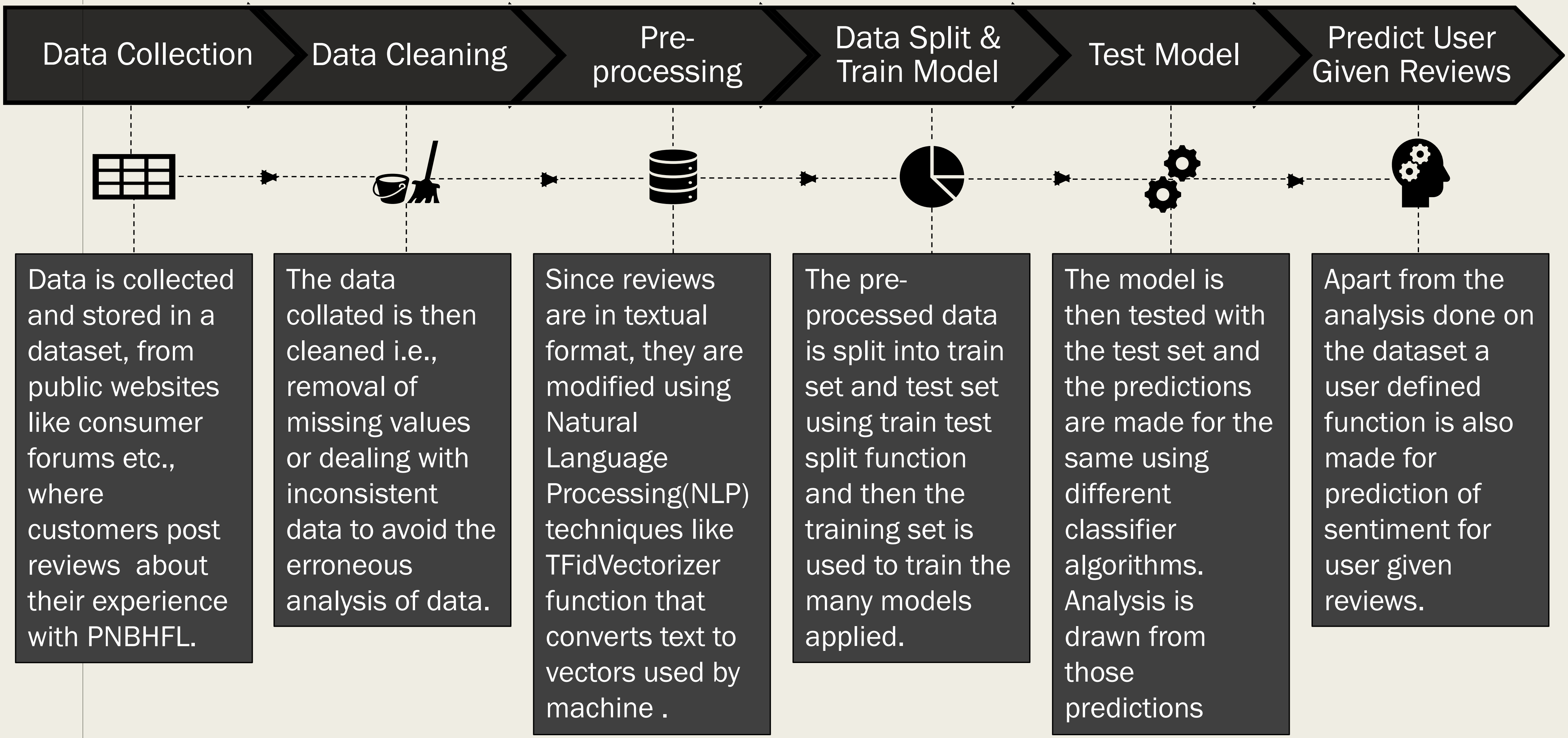
PNB Housing Finance Limited (PNB Housing) is a registered housing finance company with National Housing Bank (NHB).

- It provides housing loans to individuals and corporate bodies for purchase, construction, repair and upgradation of houses.
- It also provides loans for commercial space, loan against property and loan for purchase of residential plots

AIM 🎯

- Company customers give ratings and write reviews about the services on company websites. These reviews and ratings help housing finance companies to evaluate services provided and take necessary action to improve customer service.
- The objective of the project is to create a model to analyze customer reviews and predict customer sentiment through the reviews with the help of NLP techniques

Approach Used to Handle the Objective of This Project



Data Collection

Data Cleaning

Pre-processing

Data Split & Train Model

Test Model

Predict User Given Reviews

Data is collected and stored in a dataset, from public websites like consumer forums etc., where customers post reviews about their experience with PNBHFL.

The data collated is then cleaned i.e., removal of missing values or dealing with inconsistent data to avoid the erroneous analysis of data.

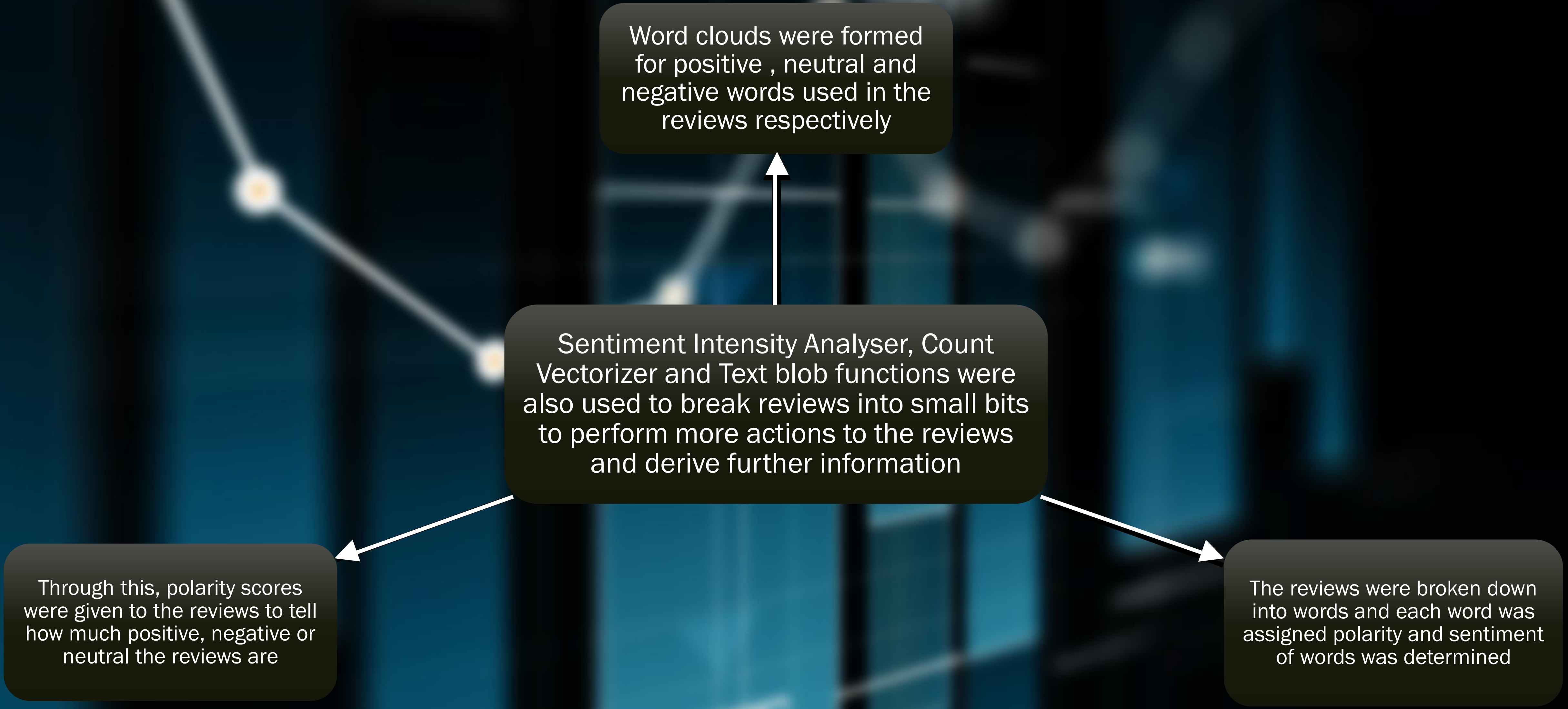
Since reviews are in textual format, they are modified using Natural Language Processing(NLP) techniques like TFidVectorizer function that converts text to vectors used by machine .

The pre-processed data is split into train set and test set using train test split function and then the training set is used to train the many models applied.

The model is then tested with the test set and the predictions are made for the same using different classifier algorithms. Analysis is drawn from those predictions

Apart from the analysis done on the dataset a user defined function is also made for prediction of sentiment for user given reviews.

The five classification algorithms applied in the project were Logistic Regression , Random Forest Classifier Naïve Bayes ,K Nearest Neighbour and lastly, Support Vector Machine .



Word clouds were formed for positive , neutral and negative words used in the reviews respectively

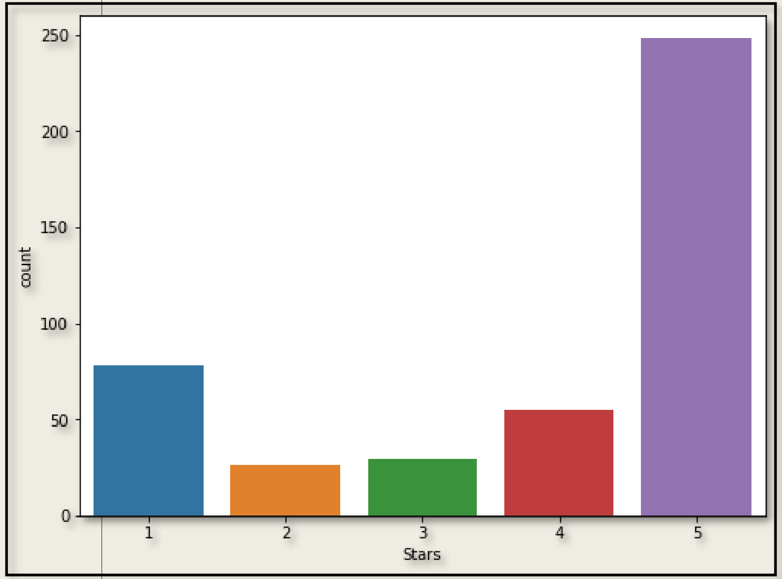
Sentiment Intensity Analyser, Count Vectorizer and Text blob functions were also used to break reviews into small bits to perform more actions to the reviews and derive further information

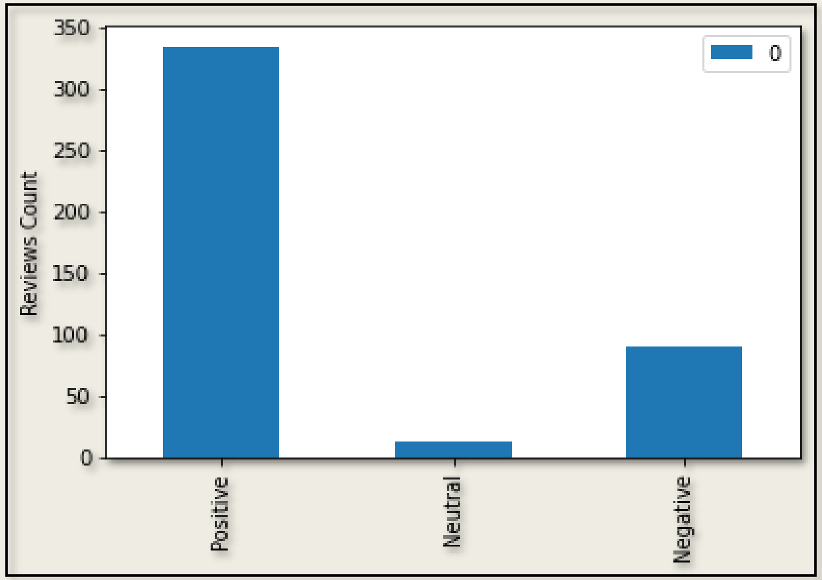
Through this, polarity scores were given to the reviews to tell how much positive, negative or neutral the reviews are

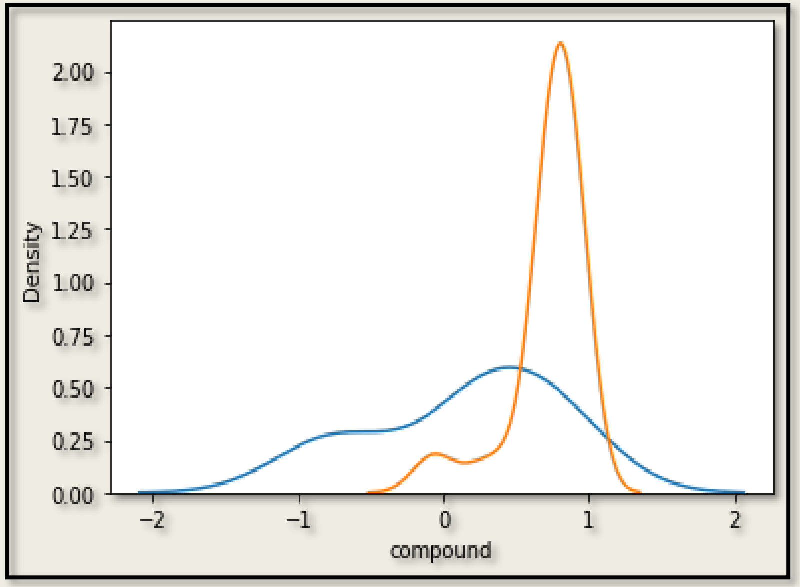
The reviews were broken down into words and each word was assigned polarity and sentiment of words was determined

Program Execution

- Exploratory Data Analysis





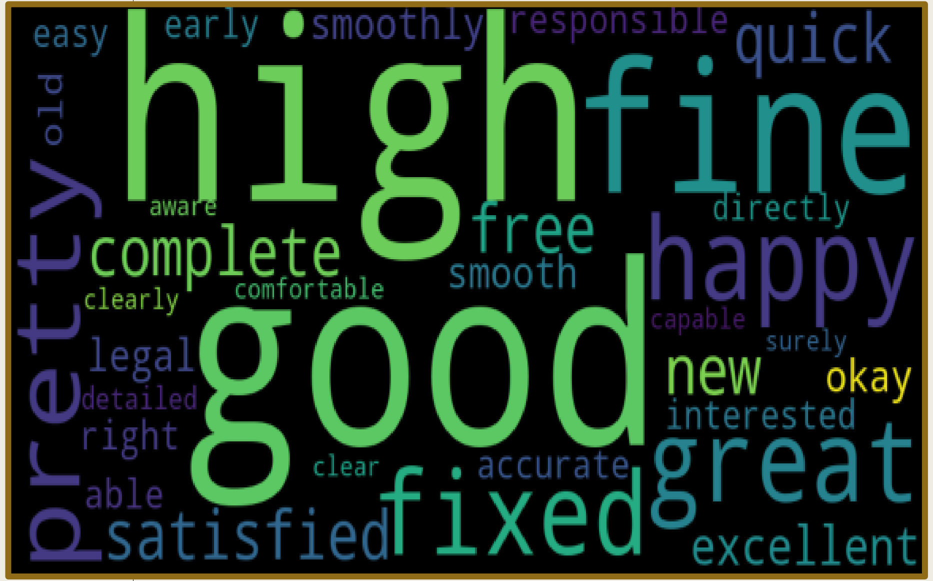


Graph for the distribution of reviews in the form of stars from the dataset.

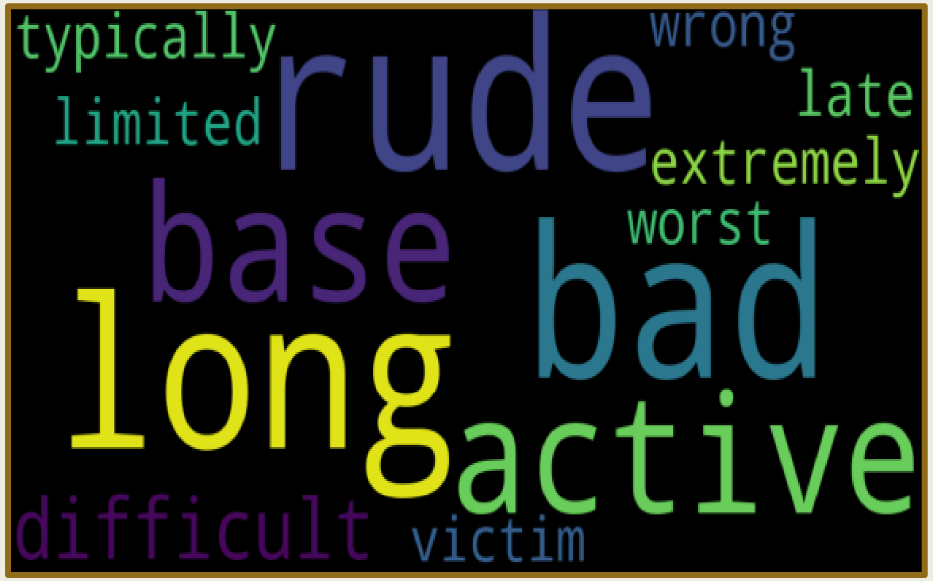
Plot of the count of different types of reviews after segregating them.

Density Plot Distribution for Positive and Negative reviews across the entire data.

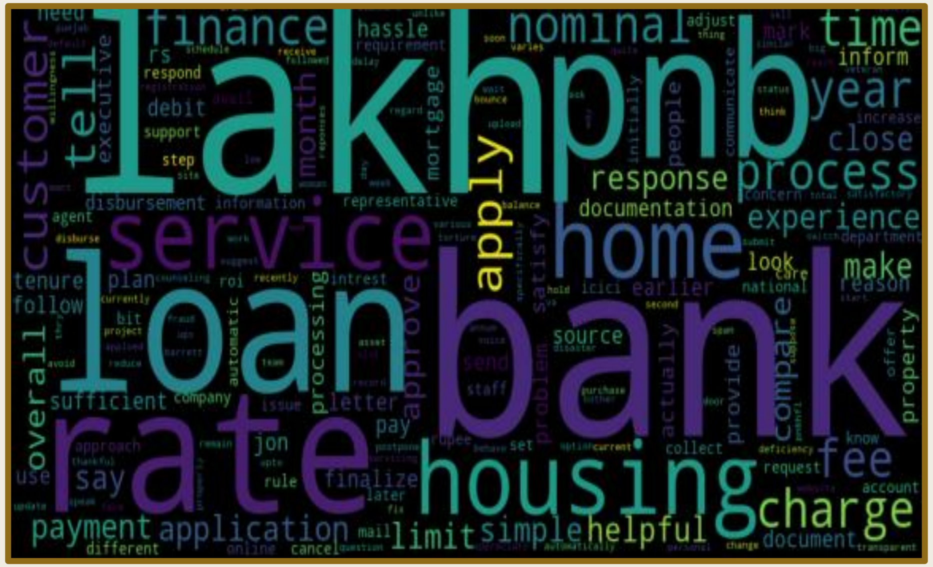
- Word Clouds For Different Types of Words



Positive Words



Negative Words



Neutral Words

• Analysis from different Classification Algorithms Applied

The following is done on data that is pre-processed using NLTK module where every review is cleaned ,tokenized and lemmatized

❖ Logistic Regression Algorithm

Fig. 1 shows the PR Curve (precision vs recall) which shows that in overall the classifier is pretty good with an average precision of 0.87 .

Fig. 2 shows the plot of ROC Curve for the model that shows the tradeoff between the true positive rate and the false positive rate of the logistic classifier and is good model in predicting the sentiment.

Fig. 3 shows the confusion matrix for the classifier that describes the performance of the classifier. The accuracy from the matrix is 0.74 with an error rate of 0.26 .

Accuracy of the model on test set is 0.76.

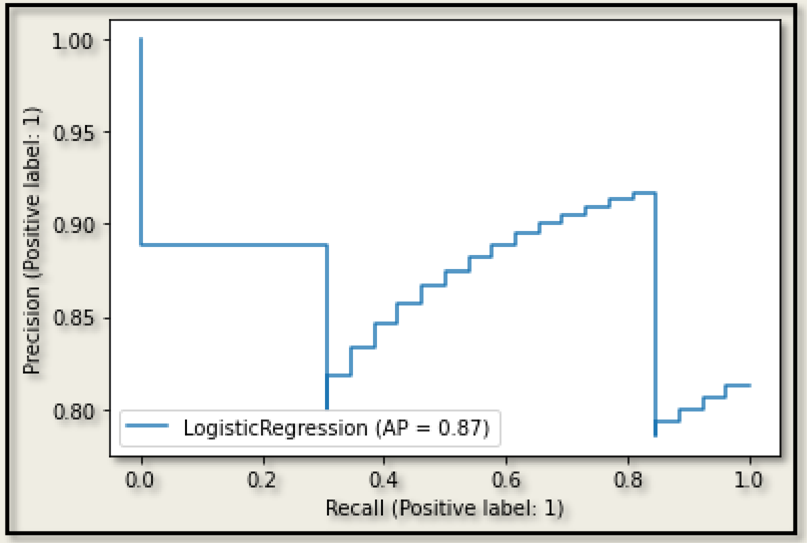


Fig. 1

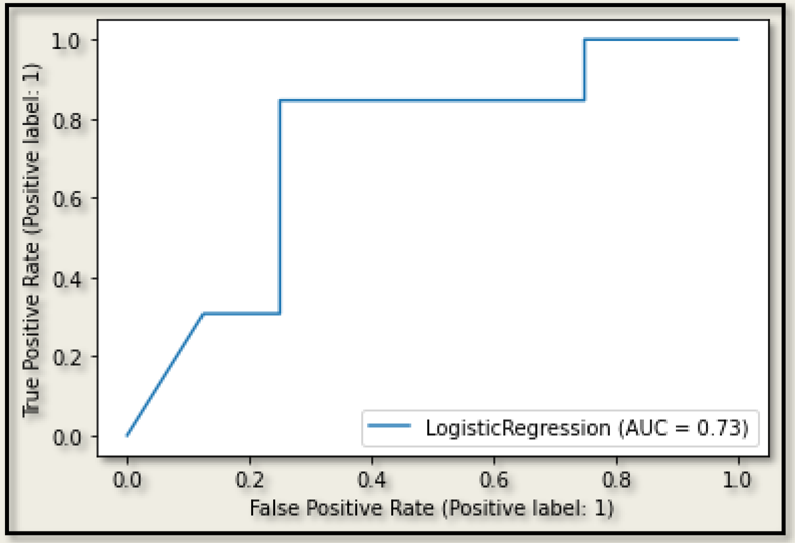


Fig. 2

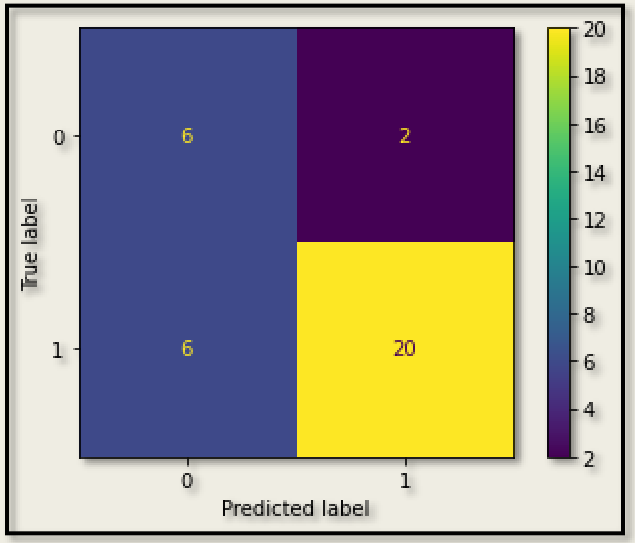


Fig. 3

❖ Random Forest Algorithm

Fig. 1 shows the PR Curve (precision vs recall) which shows that in overall the classifier is moderately good with an average precision of 0.86 giving high precision along with high recall.

Fig. 2 shows the plot of ROC Curve for the model that shows the tradeoff between the true positive rate and the false positive rate of the random forest classifier and is good model in predicting the sentiment.

Accuracy of the model on test set is 0.67

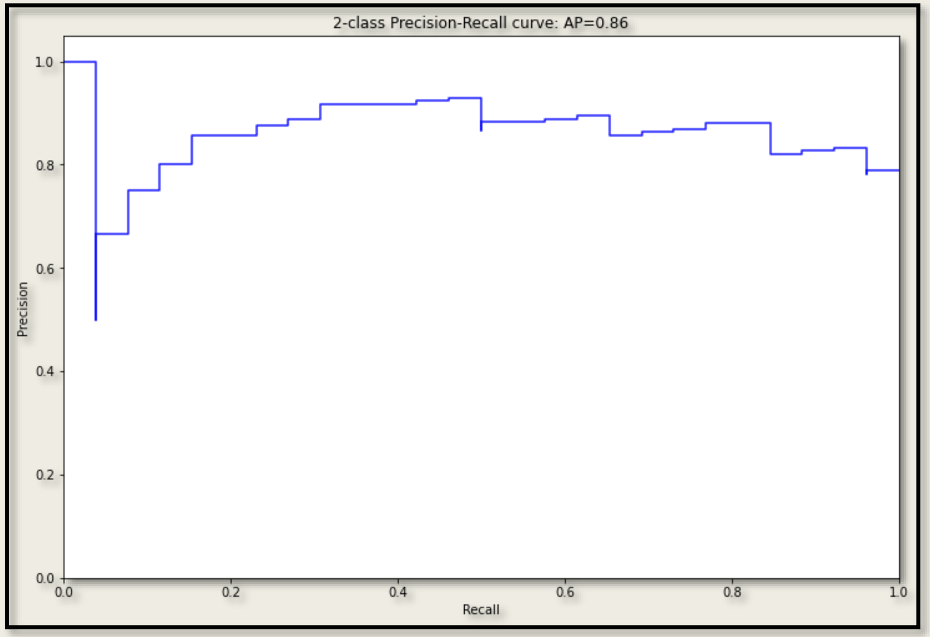


Fig. 1

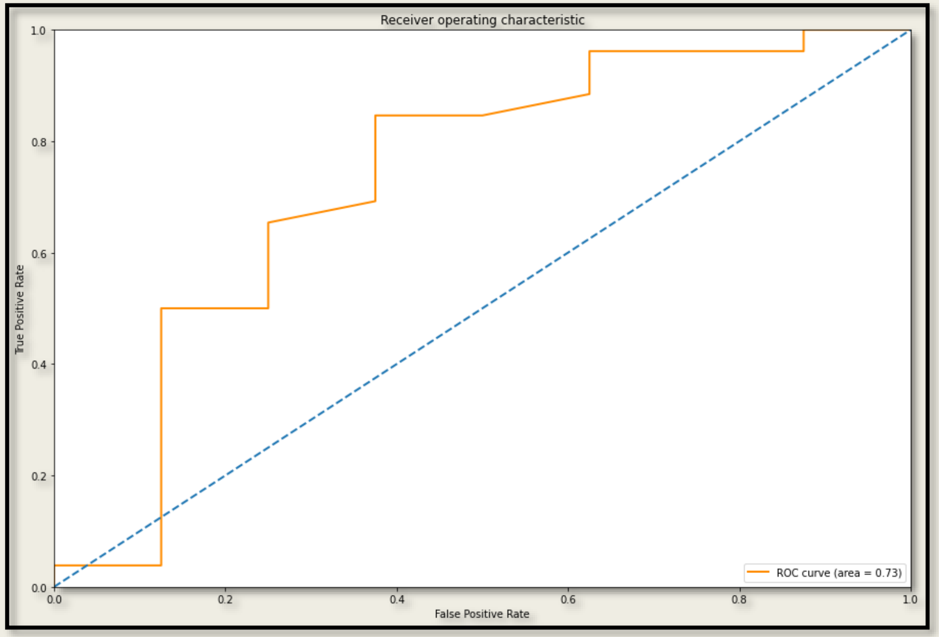


Fig. 2

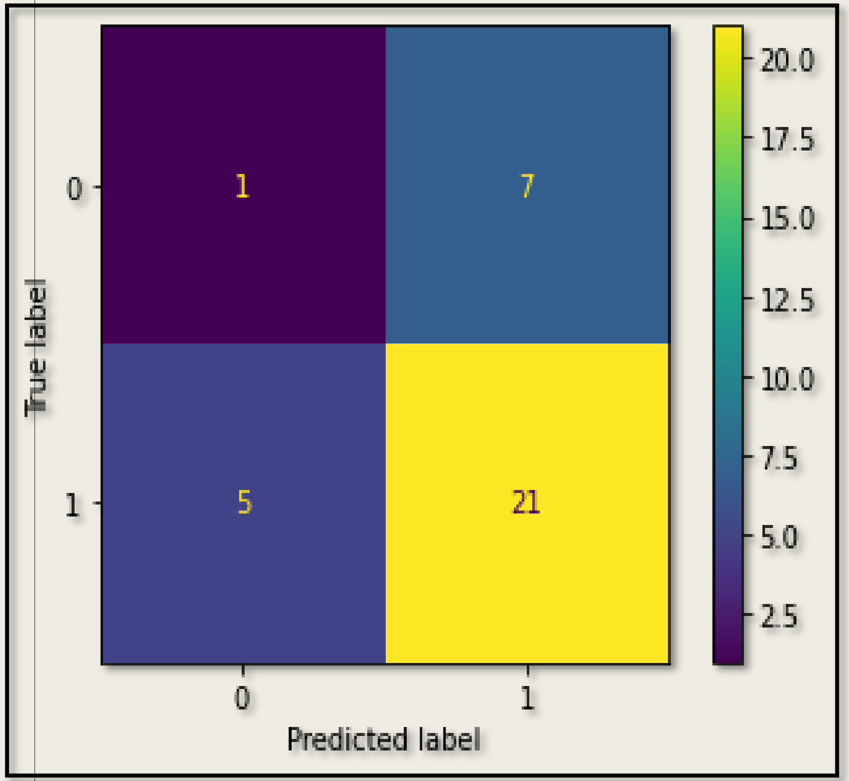


Fig. 1

❖ Naïve Bayes Algorithm

Fig. 1 shows the confusion matrix for the classifier that describes the performance of the classifier. The accuracy from the matrix is 0.65 with an error rate of 0.35 .

Accuracy of the model on test set is 0.647 .

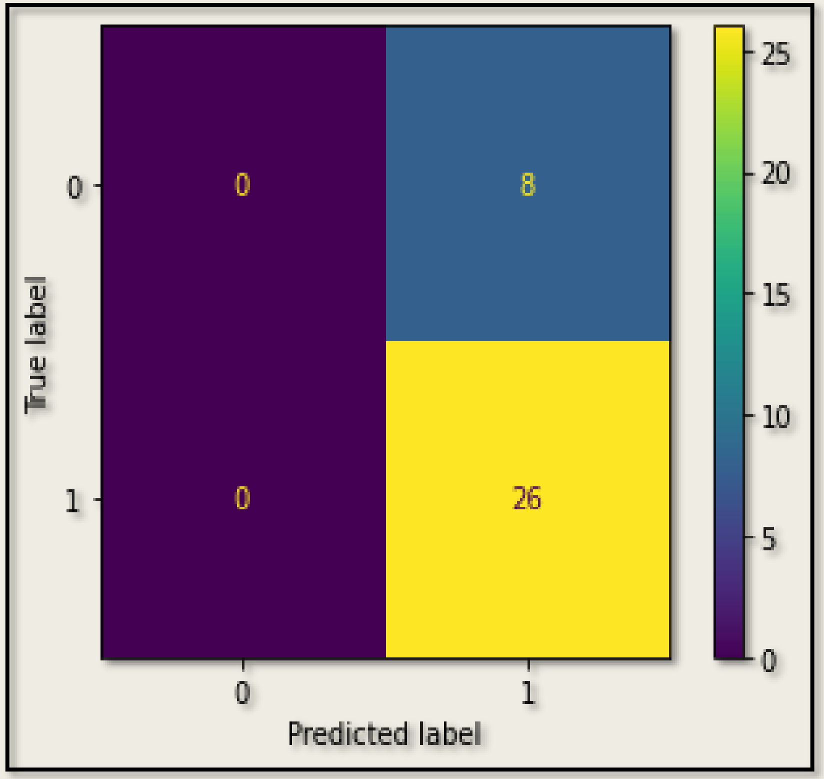


Fig. 2

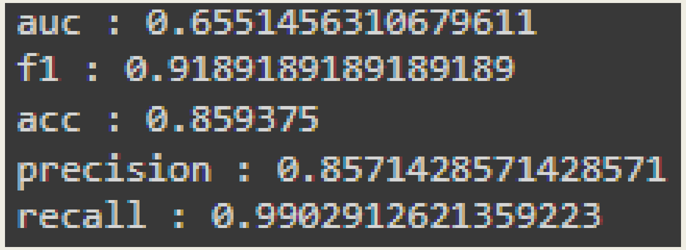
❖ K Nearest Neighbour Algorithm(KNN)

Fig. 2 shows the confusion matrix for the classifier that describes the performance of the classifier. The accuracy from the matrix is 0.74 with an error rate of 0.26 .

Accuracy of the model on test set is 0.764 .

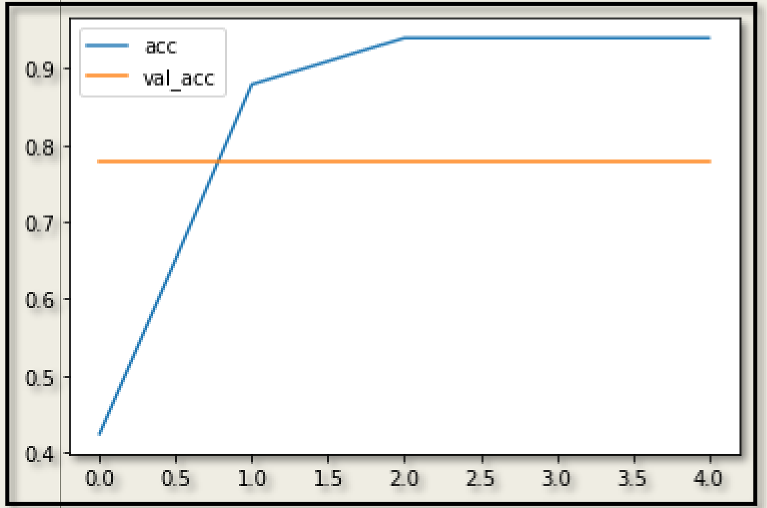
❖ Support Vector Machine Algorithm

From the figure we can see that the accuracy of the model on test set is 0.859 along with a precision of 0.857 and a recall of 0.99 .

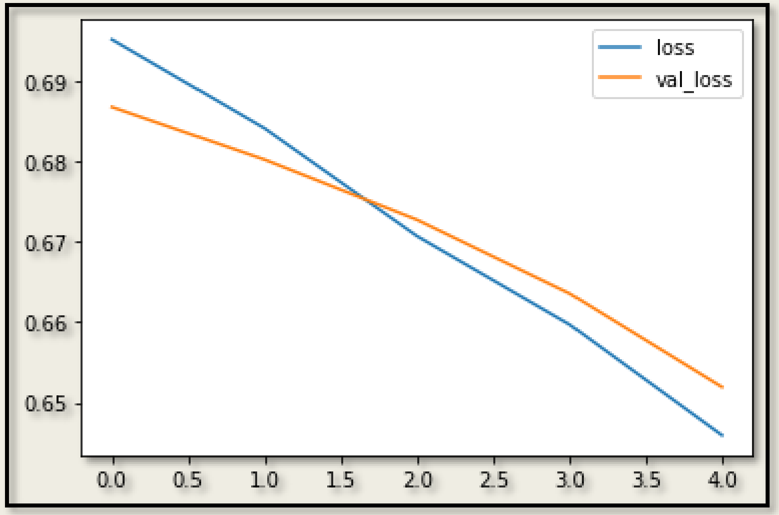


```
auc : 0.6551456310679611
f1 : 0.9189189189189189
acc : 0.859375
precision : 0.8571428571428571
recall : 0.9902912621359223
```

• Analysis of Text Classifier Model using LSTM module



This graph depicts the accuracy achieved on the train dataset(in blue) and the accuracy on the test set (in orange) . The accuracy obtained on the train set(0.939) is more than that of the test set(0.777).

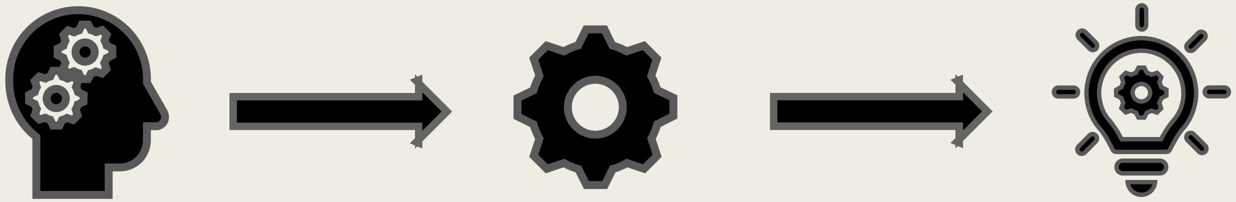


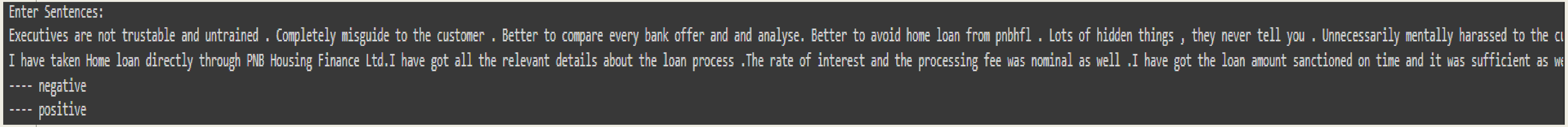
We can see the plot of loss incurred through the model on the train dataset (in blue) and on the test set (in orange). The loss on the test increases gradually .The loss on the test set(0.65) is slightly more than on train set(0.64) .

• Prediction of Sentiment for User Given Reviews

Other than just algorithm application this project also includes prediction of sentiment of reviews given by the user apart from just the dataset . As we have seen that Support Vector Machine Algorithm has the highest accuracy on the test set among all the algorithms applied. Therefore we will use it on user given review as well to help give us a idea if the review is a positive sentiment or a negative sentiment.

We can see a sample below for the above said conclusive statement .





```
Enter Sentences:
Executives are not trustable and untrained . Completely misguide to the customer . Better to compare every bank offer and and analyse. Better to avoid home loan from pnbhfl . Lots of hidden things , they never tell you . Unnecessarily mentally harassed to the cu
I have taken Home loan directly through PNB Housing Finance Ltd.I have got all the relevant details about the loan process .The rate of interest and the processing fee was nominal as well .I have got the loan amount sanctioned on time and it was sufficient as we
---- negative
---- positive
```

Conclusive Outcome

The most important thing is to be able to extract the relevant features from the raw source of data so that it helps in building highly strategic decisions and increase the predictive power of the models.

As and when we have more balanced the results and the model performance becomes more and more better.

We have seen that SVM model has the highest accuracy of 0.859 hence is used for prediction for user given reviews as well.

Furthermore modifications/

additions can be done to this project to suit the requirements of the user.

THANK YOU!