STACK OVERFLOW QUESTION RATING CLASSIFICATION

Team:
Deshmukh, Prathmesh
Lewis, Gavin Henry
Katapally, Manognya

Data Mining (CSCI-B 565 Spring 2022)

PRESENTATION OUTLINE

- Introduction
- Exploratory Data Analysis
- Data Cleaning
- Implementation of Classification Model
- Results
- Conclusion & Future Work

INTRODUCTION

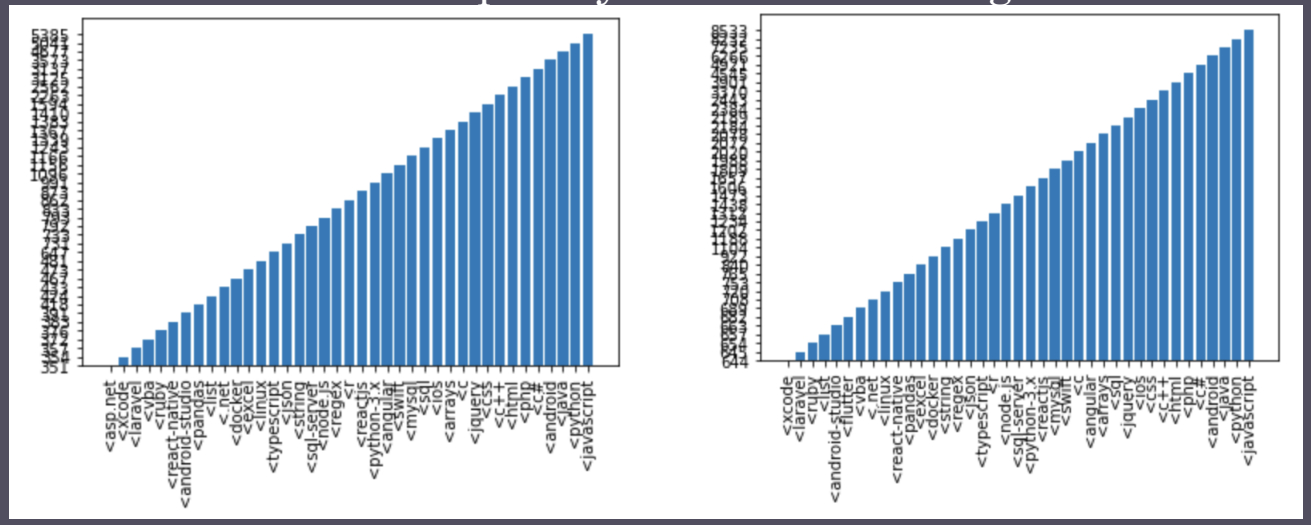
The project covers classification on the “60k Stack Overflow Questions with Quality Rating” data set that is available on Kaggle.

Full network pre-trained model – ALBERT to implement the classification model.

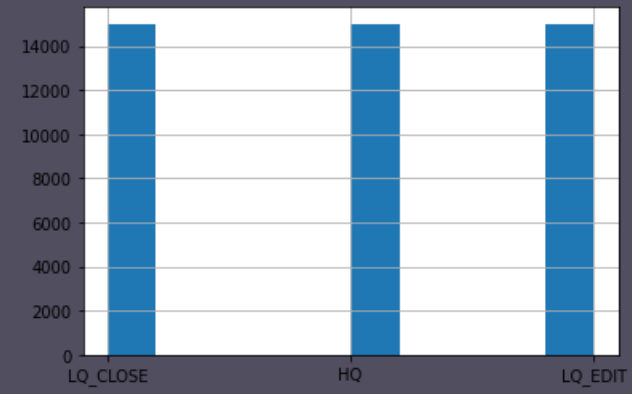
Projecting the data to a lower dimension to preform clustering using TSNE.

EXPLORATORY DATA ANALYSIS

- Explored the relation between the 'Tags' attribute and the target class 'Y'
- Visualized the frequency of individual 'tag' elements.

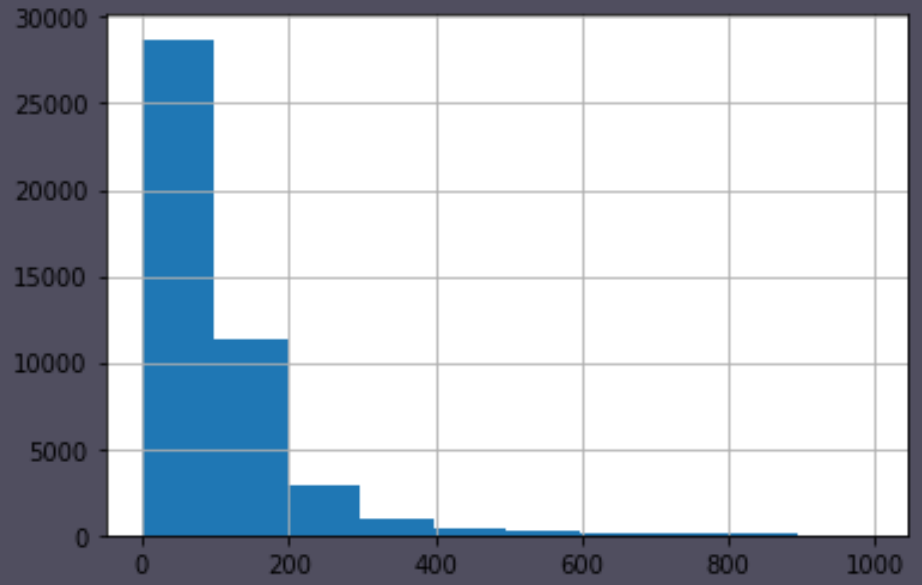


- Visualized the distribution of the Target attribute 'Y'. The attribute is thus balanced.



- Plotted the histogram for the length of the body string removing the outliers in the data.
- The following is the distribution of the length of the body string.

```
count    45000.000000
mean       107.376200
std        120.708652
min          0.000000
25%         46.000000
50%         76.000000
75%        128.000000
max       5412.000000
Name: len_body, dtype: float64
```



- Since the mean is 107, we have taken the embedding size as 128.

DATA CLEANING

- Explored the usage of two separate approaches for data cleaning:
  - Html2text: This approach was initially used to check and remove the unwanted html tags in the 'body' attribute in the data.
  - Using Regular Expression and Contraction: This approach filters the tags as well as removes the shortened words in the data.
- We are using the second approach as it provides us with better results.

IMPLEMENTATION

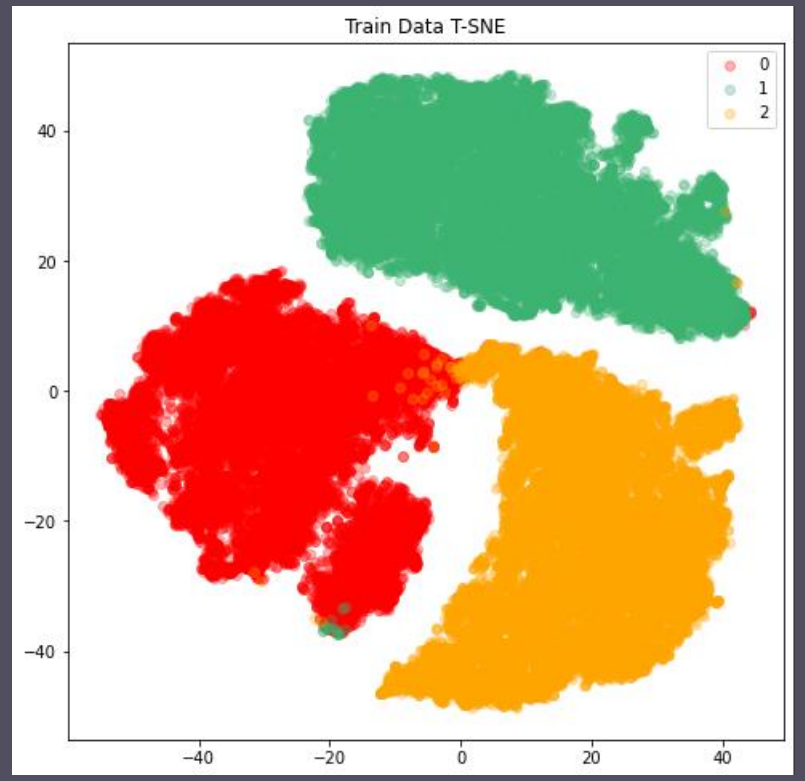
- We have tried the implementation of two models:
  - BERT
  - ALBERT
- We have decided to use the ALBERT model as it is computationally more optimal to use this.
- The ALBERT model has far much lower trainable parameters and thus would train faster.

RESULTS

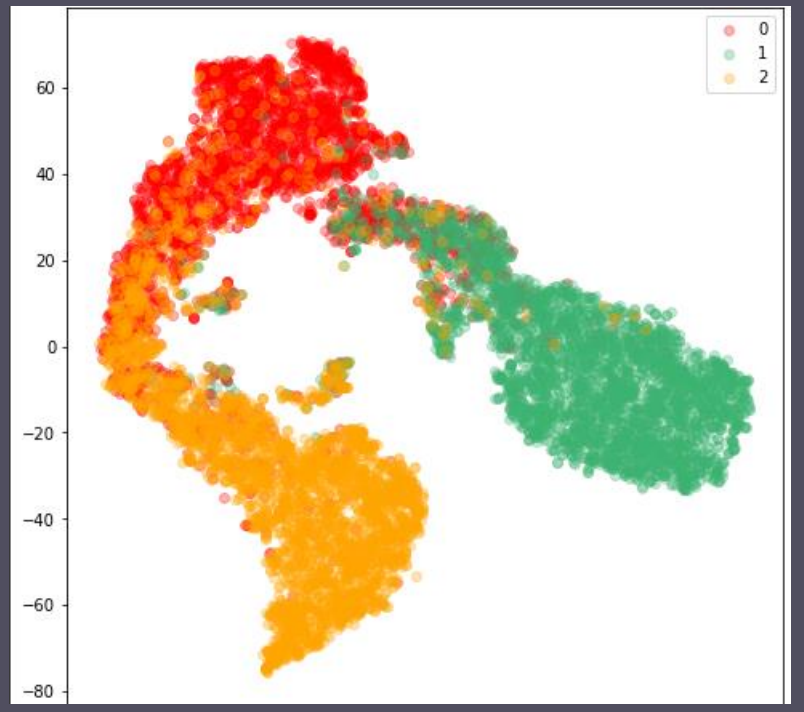
- ALBERT model performed best on the configuration of 20 epoch for training with learning rate of 3e-5 and weight decay of 5e-6.
- The model had an average loss of 0.0201 on the train data and a minimum loss of 0.3538 on the test data.
- Best model is chosen using the metric of minimum loss on the test data.

Class	Model Evaluation on Test	Accuracy
LQ_CLOSE	2686/3340	0.8042
LQ_EDIT	3065/3335	0.9190
HQ	2980/3375	0.8830
Total Accuracy on Test		0.8688

TSNE Plots on the data



Test Data TSNE plot



CONCLUSION & FUTURE WORK

- The pre-trained models have prior knowledge of English language and thus they perform better when trained on the 45000 rows in the given dataset.
- We tried implementation of BERT but due to GPU/TPU memory limitation and longer training times we opted for ALBERT model.
- The ALBERT model is efficient for producing results quickly and with training time as it has 11M trainable parameters.
- We can explore further implementation of BERT and GPT2 models on the dataset which have trainable parameters of 110M and 1.5B respectively.