Assignment 1

Shubham Nemani :- 203050011
Pranjal Saini :- 203050014
Harsh Peswani :- 203050043

Guide: Dr. Pushpak Bhattacharyya

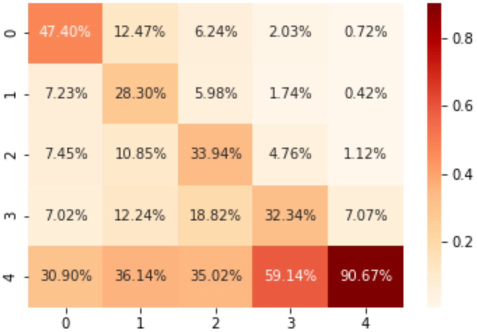
Indian Institute of Technology Bombay

February 10, 2021

Approach

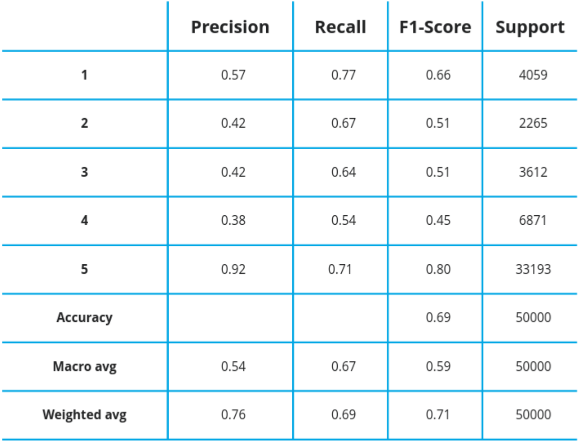
- We first performed data cleaning(convert corpus to lower case, removed punctuation, converted contraction, performed tokenization, padding)
- We used pytorch to implement and train over neural network.
- We used sklearn to report precision, recall and F1-score.
- We also used nltk library for some data cleaning tasks.
- We used pre-trained glove embedding (200d).

Confusion Matrix For Train Data



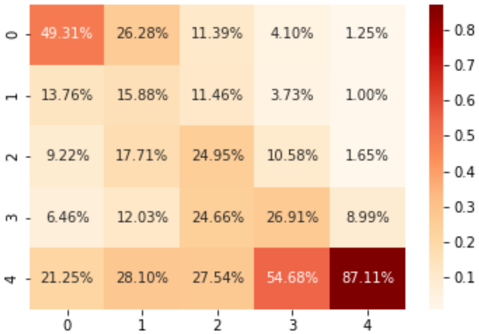
	0	1	2	3	4
0	47.40%	12.47%	6.24%	2.03%	0.72%
1	7.23%	28.30%	5.98%	1.74%	0.42%
2	7.45%	10.85%	33.94%	4.76%	1.12%
3	7.02%	12.24%	18.82%	32.34%	7.07%
4	30.90%	36.14%	35.02%	59.14%	90.67%

Precision, Recall and F1 Scores For Train Data



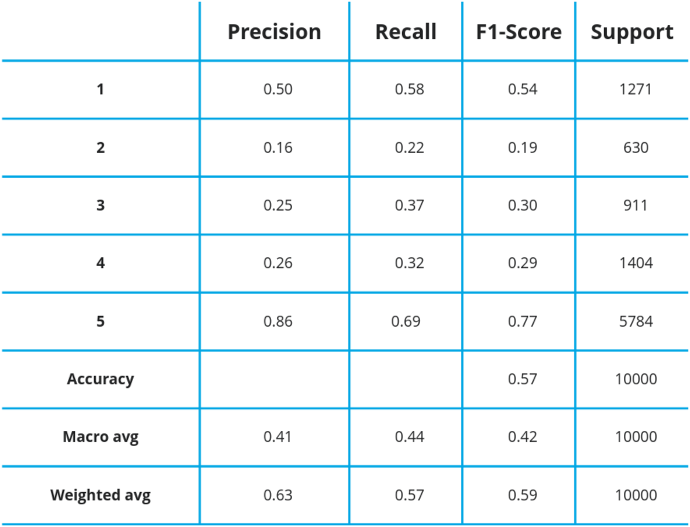
	Precision	Recall	F1-Score	Support
1	0.57	0.77	0.66	4059
2	0.42	0.67	0.51	2265
3	0.42	0.64	0.51	3612
4	0.38	0.54	0.45	6871
5	0.92	0.71	0.80	33193
Accuracy			0.69	50000
Macro avg	0.54	0.67	0.59	50000
Weighted avg	0.76	0.69	0.71	50000

Confusion Matrix For Test Data



	0	1	2	3	4
0	49.31%	26.28%	11.39%	4.10%	1.25%
1	13.76%	15.88%	11.46%	3.73%	1.00%
2	9.22%	17.71%	24.95%	10.58%	1.65%
3	6.46%	12.03%	24.66%	26.91%	8.99%
4	21.25%	28.10%	27.54%	54.68%	87.11%

Precision, Recall and F1 Scores For Test Data



	Precision	Recall	F1-Score	Support
1	0.50	0.58	0.54	1271
2	0.16	0.22	0.19	630
3	0.25	0.37	0.30	911
4	0.26	0.32	0.29	1404
5	0.86	0.69	0.77	5784
Accuracy			0.57	10000
Macro avg	0.41	0.44	0.42	10000
Weighted avg	0.63	0.57	0.59	10000

Experiment and Analysis

- **DataSet** :- Train - 40K ,Validation - 10K ,Test - 10K.
- Dataset was highly skewed towards class 5 , around 65% training examples were of class 5 , so we used **Weighted Sampling**.
- We also observed that varying validation set size affected test accuracy to some extent. It was highest when using 20% data as validation set resulting in highest test accuracy i.e. 57%.
- **Training**:
  BatchSize = 32
  Epochs = 15
  Preprocessing Time = 188 sec
  Training Time = 47.43 sec

- Accuracy and F1 score for classes decreases because of removing stop words.

  Ex : I like to play cricket . Predicted Rating - 5
  I do not like to play cricket. Predicted Rating - 5

  Since in second sentence removing stop words removed **do** and **not** so it's rating predicted 5 because of word **like**

Class	1	2	3	4	5	Overall
With Stop words	0.54	0.19	0.30	0.29	0.77	0.42
After Removing Stop words	0.16	0.12	0.13	0.08	0.72	0.24

Table: F1 score for test Data

Thank You