CS-772: Assignment #4

Team Id : 25
Debabrata Biswal, 203050024
Ashish Aggarwal, 203050015
Keshav Agarwal, 203050039

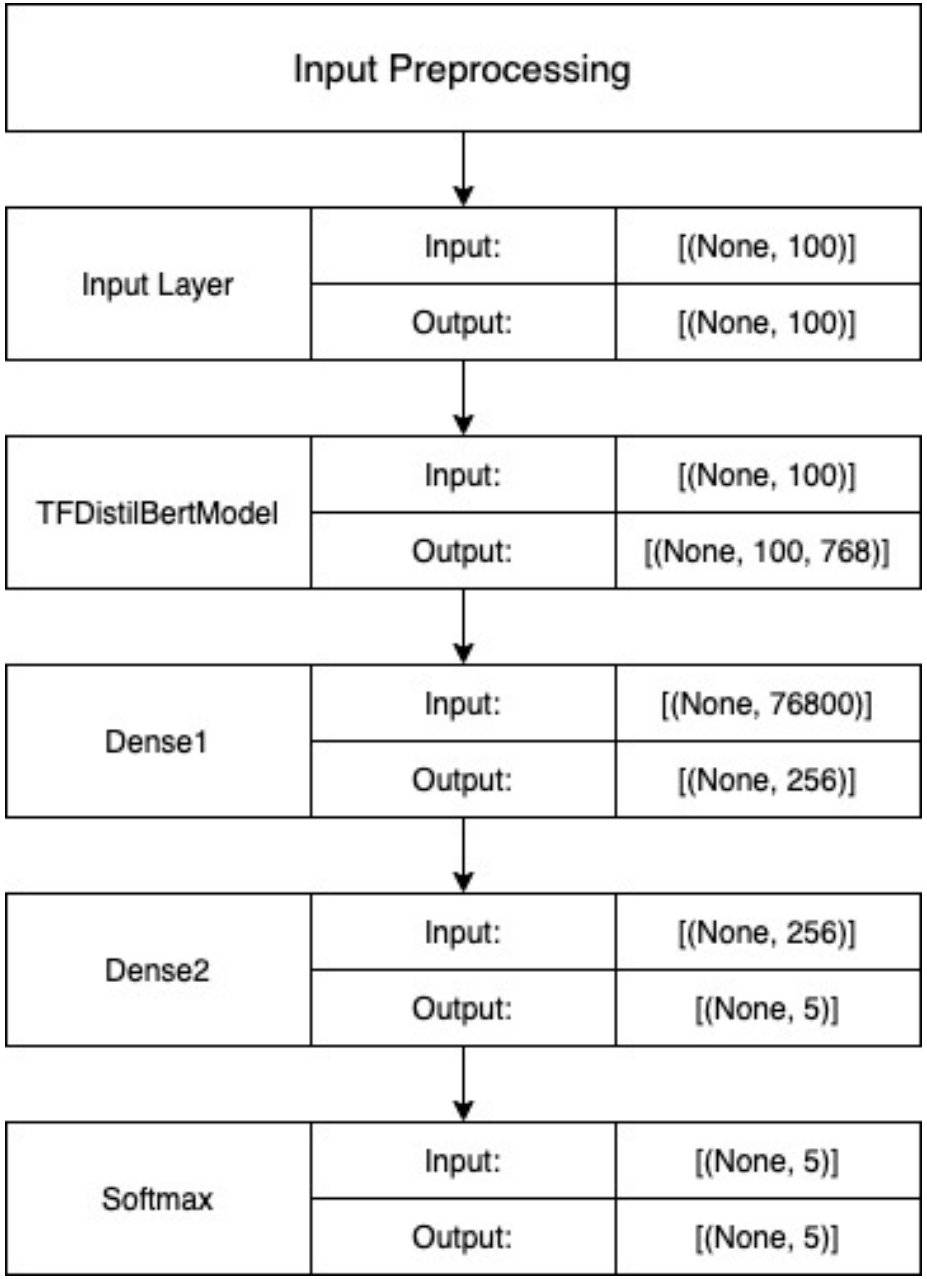
Problem statement

- Identifying the rating of a review using sentiment analysis
- Input dataset contains reviews and their rating score between 1 to 5 - low to high
- Given a review, we predict the rating score.

Architecture

Libraries

Data preprocessing - scikit-learn for validation split - transformers library
Transformer - TFDistilBertModel
Neural Network - TensorFlow – keras
Prediction and Evaluation - keras - scikit-learn
Explainability - LIME



Results

Model	Precision	Recall	F1-Score	Accuracy
RNN	0.66	0.60	0.62	0.60
LSTM	0.67	0.61	0.63	0.61
Bi-LSTM	0.69	0.61	0.64	0.61
GRU	0.68	0.62	0.64	0.62
Bi-GRU	0.69	0.62	0.64	0.62
Transformer	0.72	0.71	0.72	0.71

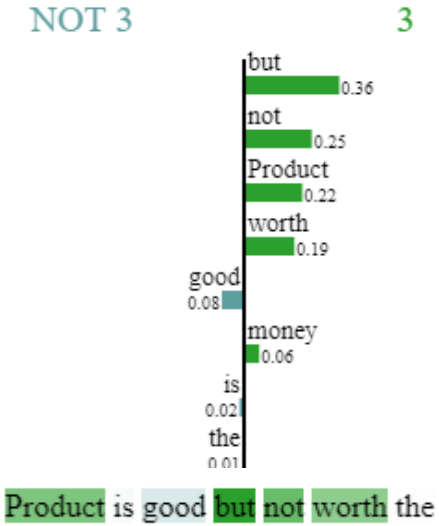
- With balanced data, Transformer has overall best accuracy
- With Unbalanced data, Transformer accuracy goes to 76%

Qualitative analysis with some examples

'good' word unigram count per class		27	57	82	348	504
Review	**Rating**	**Class 1**	**Class 2**	**Class 3**	**Class 4**	**Class 5**
So far so good. Everything is as described and worked great the first time.	5	0.0	0.0	0.0	0.497	0.503
Product is not good.	1	0.931	0.001	0.068	0.0	0.0
Product is not good but too cheap.	2	0.066	0.934	0.0	0.0	0.0
Product is good but not worth the money.	3	0.0	0.0	0.952	0.047	0.0
Product is good.	4	0.0	0.0	0.003	0.989	0.008
Product is good and worth the money.	5	0.0	0.0	0.0	0.091	0.909

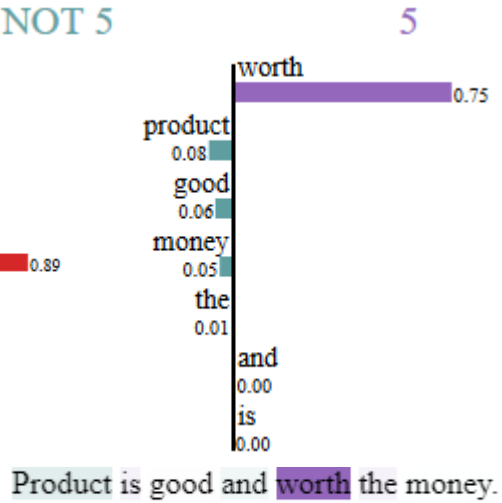
Product is not good.

Product is not good but too cheap.



Product is good but not worth the money.

Product is good.



Product is good and worth the money.

Confusion matrix

	1	2	3	4	5
1	815	238	122	38	58
2	153	253	146	46	32
3	72	115	402	219	103
4	13	32	164	711	484
5	32	29	74	740	4909

Confusion Matrix

	1	2	3	4	5
1	26942	6628	15758	18611	22653
2	6628	7232	6550	6576	6792
3	15758	6550	20069	17286	17875
4	18611	6576	17286	28936	26743
5	22653	6792	15875	26743	83040

Unigram Confusion Matrix

- High confusion between rating 4 and 5 due to many overlapping words
- Adjacent classes have high overlapping unigrams
- Unigrams for rating 3 are spread over all the classes
- Diagonal entries in the Unigram Confusion Matrix shows the word count for each class

Sentiment Analysis Demo

Sentiment Analysis

Enter review

Analyse

Review	Rating	Class 1	Class 2	Class 3	Class 4	Class 5
This camera is awesome, I will not recommend this to anybody	1	0.997	0.0	0.0	0.0	0.003