Code Mixed Sentiment Analysis

Submitted by: Jatin Dholakia (16110066)

GitHub link: https://github.com/JatinDholakia/NLP-Assignments/blob/master/Assignment3-16110066.ipynb

Pre-processing

- Removal of links and tags.
- Replacing all non-ascii characters with white spaces.
- Converting all text to lowercase.
- Converting of all characters to their index in vocabulary.
- Padding all tweet sequences to equal length of 150.

Model

The model is a character level sentiment classifier model. Hence the vocabulary consists of all the lowercase characters from a-z, padding token (PAD), unknown token (UNK) and space(‘ ‘).

Batch size = 128.

Embedding size = 128

Epochs = 25

The following model is used:

The first layer is an Embedding layer. Output dimension = (150,128)

1-D Convolution layer. Output dimension (146,128)

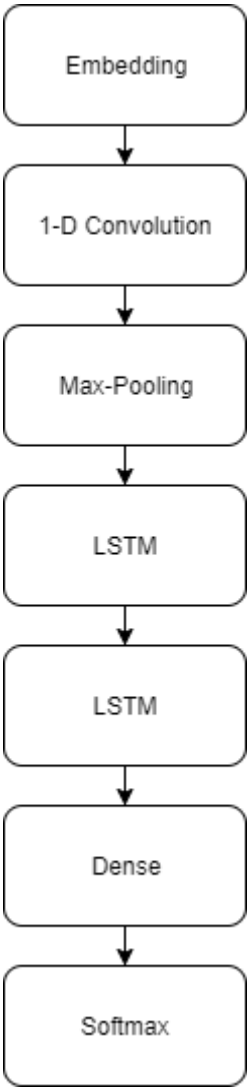
Max pooling layer. Output dimension = (48,128)

LSTM. Output dimension = (48,128)

LSTM. Output dimension = (128,)

Dense layer. Output dimension = (3,)

SoftMax activation layer. Output dimension = (3,).



The model was trained for 25 epochs. The scores obtained are given below.

Evaluation

Overall accuracy = 0.45050829%

Metric	Negative	Neutral	Positive
Recall	0.45778612	0.61803714	0.22680412
Precision	0.44283122	0.44721689	0.47826087
F-score	0.4501845	0.51893096	0.30769231

References:

- Aditya Joshi, Ameya Prabhu, Manish Shrivastava and Vasudeva Varma. Towards Sub-Word Level Compositions for Sentiment Analysis of Hindi-English Code Mixed Text.
- NLP-character-embeddings - https://towardsdatascience.com/besides-word-embedding-why-you-need-to-know-character-embedding-6096a34a3b10