

Sentiment Analysis with different models and Their Comparison:

The submitted text classification method is the end result of comparison of two separate NLP models and choosing the best one. The first is a simple LSTM based model which needs some preprocessing. The data preprocessing phase includes Unwanted text removing, Decode contractions, Punctuation removing, Stopword removing and One-hot encoding(in alphabetical order) of the words in each sentences. The LSTM model which is a sequential one includes an embedding layer, LSTM layer, dense layers, and dropout layers. There is only one LSTM which has 300 neurons followed by a dense layer with 20 neurons, kernel_regularizer L2(1e-6) and ReLU activation with dropout of 0.3 followed by another dense layer with 20 neurons, kernel_regularizer L2(1e-6) and ReLU activation. Output layer has 5 neurons and softmax activation function. The model is the result of multiple trial and error, and some research from internet . It is then trained with ADAM optimizer with learning rate 0.02, considering the categorical crossentropy loss.

The second model is a Transformer based model with attention mask. Here BERT autotokenizer is used with trainable input mask. Before that preprocessing is done involving unwanted text removing, contractions decoding. The model involves an embedding layer based on BERT autotokenizer then an attention mask followed by a global maxpool layer. Followed by a dense layer with 128 neurons and ReLU activation with 0.1 dropout with another dense layer with 32 neurons and ReLU as activation function. The final output layer has 5 neurons with softmax activation function. Finally the BERT parameters are set as trainable for fine tuning. This model is trained with ADAM optimizer with learning rate 5e-5 and epsilon 1e-8 considering the categorical crossentropy. A SpaCy-'en_core_web_sm' model is used as benchmark comparison.

Model Name	Accuracy Score
LSTM	68.79
SpaCy- en_core_web_sm	71.39
BERT	88.03