

biLSTM with Cellula toxic data

- clean the data set
 - Converts all characters to lowercase.
 - Removes all special characters and punctuation using regex
 - replacing multiple spaces with a single space
 - Splits the text into individual words
 - Removes stop words
 - add lemmatization
- Apply word tokenization
- split the data to train and test
- vectorized the data, assigns each unique word an integer index
- all the sequences with the same length and apply padding
- apply label encoder on the target
- build bilstm model
 - image description as an input for the first bilstm
 - query as an input for the second bilstm
 - concatenate them
 - add two dense layers with relu
 - output layer with softmax for 9 classes
- add two callbacksmodel check point and early stooping
- fit the model

```
Epoch 1/5
59/60 ————— 0s 34ms/step - accuracy: 0.9963 - loss: 0.0147
Epoch 1: val_loss improved from inf to 0.30765, saving model to best_bilstm_model.keras
60/60 ————— 2s 37ms/step - accuracy: 0.9963 - loss: 0.0146 - val_accuracy: 0.9646 - val_loss: 0.3076
Epoch 2/5
59/60 ————— 0s 33ms/step - accuracy: 0.9961 - loss: 0.0161
Epoch 2: val_loss improved from 0.30765 to 0.29811, saving model to best_bilstm_model.keras
60/60 ————— 2s 36ms/step - accuracy: 0.9961 - loss: 0.0161 - val_accuracy: 0.9688 - val_loss: 0.2981
Epoch 3/5
58/60 ————— 0s 34ms/step - accuracy: 0.9967 - loss: 0.0081
Epoch 3: val_loss did not improve from 0.29811
60/60 ————— 2s 36ms/step - accuracy: 0.9968 - loss: 0.0080 - val_accuracy: 0.9688 - val_loss: 0.3368
Epoch 4/5
60/60 ————— 0s 34ms/step - accuracy: 0.9949 - loss: 0.0116
Epoch 4: val_loss did not improve from 0.29811
60/60 ————— 2s 36ms/step - accuracy: 0.9949 - loss: 0.0116 - val_accuracy: 0.9646 - val_loss: 0.3131
Epoch 4: early stopping
Restoring model weights from the end of the best epoch: 2.
```

- after prediction the weighted f1 score 0.9555

