Emotion Classification in Hindi Language

The dataset is having 517 examples of 4 emotions i.e. happy (150), sad (103), angry (129) and neutral (127) written in Hindi language. I am using google colab as a working environment to manipulate dataset as well as build, train and test the classifier for the same.

First I divided all the sentences into train and test set with 80% train set and 20% test set.

I extracted the words from all the sentences i.e. I tokenized the sentences.

Then I found the frequency of all the tokens.

After finding the frequency I found that the data contained unnecessary words i.e. the words which doesn't give any relation with sentiment whatsoever.

So I used the CNLTK library to remove the stop words from the token list.

Then I found the frequency of all the remaining tokens.

After finding the frequency I decided to convert every sentence into the one hot encoding of the remaining tokens i.e. if a word in a sentence is part of the remtoken list than that one hot encoding is taken and so on for all the words of a single sentence.

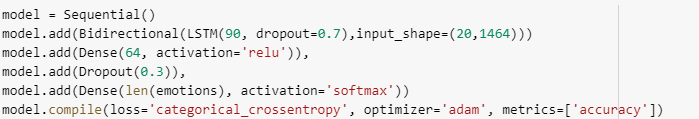
Now each one hot encoding list of each sentence are not of same size so I padded each sentence one hot encoding list with the extra one hot vectors to make all the list of all the sentences of equal length.

The maximum number of tokens in a sentence is 20 so converted all other lists to 20 size.

I also removed those sentences which didn't contain any word from the remtoken list.

So now the final Training set and testing set is ready to feed into the model.

The model is described below:



Training Dataset: 396 samples

Validation Dataset: 171 samples

Testing Dataset: 143 samples

Batch Size: 64

No of Epochs: 150

I divided the Training set into Training and Validation set with 70-30 split as the dataset is very small so it was over fitting on the training dataset.

After training the model it is giving 89% accuracy on Training set and 61% accuracy on the validation set.

After that I tested the model and gave nearly 60% accuracy on the test set.

After that I also tested individual sentences of all the emotions of which it could give the answers perfectly.

Conclusion:

The dataset is too small for using the Recurrent Neural Networks.

We can increase the accuracy of the model by introducing more examples.

We can also use pre trained model on the given small  dataset so that we can achieve more accuracy without increasing the dataset i.e. we can do transfer learning by freezing the existing layers of the existing model and adding our desired layers at the end of this model and training on those levels.

For testing we would unfreeze already trained model so that we can perform testing on the whole model.

This is also a promising technique for performing classification on small dataset efficiency.