Abhinav Gupta
abhinavg@usc.edu

CAIS++ Pre-Semester Project

Type of Task: Classification

After spending a significant amount of time thinking about the problem, I chose to do a classification task because even though the level of emotions seems to be sequential, and could be given a continuous value, I chose to approach the problem as classifying 3 different emotions which can be considered as 3 discrete classes. This would mean that the task is a classification task, and the algorithm will try to identify which tweet belongs to which class.

Data preprocessing

- Shuffling the data set: I shuffled the data so that it's randomly trained, and the order doesn't influence the training
- Creating a word bag / vocabulary: I created a word bag for all the words so that they can be assigned indices.
- Padding the inputs and converting to fixed size vectors: Since the input size can vary as tweets can be of any number of words, I then converted the input to vectors with the indices of the words from the word bag, and then padded it to a maximum of 140 words (since twitter only allows 280 characters at maximum and each word with one character and one space between words would amount to 140 maximum possible words).
- Encoding the labels : The labels were also converted to a vector and I used one hot encoding for that purpose.

Model

I used an RNN model with LTSM layers for this task. Relu activation function was used in most layers as it is the most widely used function for tasks like these and a sigmoid function was used at the last layer to predict one of the different classes for the model.
The data was split into training, validation and testing with a 60%, 20% and 20% split. The optimizer used was RMSprop, and the loss was binary crossentropy. Evaluation metrics used were binary accuracy.