Hilari Fan
hilarifa@usc.edu

2. Indicate whether you chose to do a classification or regression task and why.

I decided to do a classification task since the possible valence values were discrete: 0, 2, and 4. Therefore, I didn't choose to do a regression task whereby the predicted output would be a continuous value.

3. Explain any data preprocessing you tried and used. If you used an external word embedding model to preprocess your tweets, please provide a name and reference to it.

I used the GloVe word embedding as used in the RNN notebook and pandas in order to separate the tweets and labels from the csv data file.

4. What were the features (post preprocessing) you trained the machine learning model on, and what were the targets (post preprocessing)? Please list the feature names, domain, and type in a table (similar to the one we provided in the Dataset section.

Name	Domain	Type
Tweet (feature)	string	string
Valence (target)	{0, 2, 4}	int

5. Indicate the type of machine learning model you used and any hyperparameters you adjusted or specified. If you used a neural network, please list the input, hidden, and output layers you used with their respective sizes and activation functions. Why did you settle on this model/architecture? What were the steps that led you to it?

I used a recurrent neural network for a sequential model (similar to the RNN notebook). The model has a pre-trained embedding layer, two LSTM layers, and 2 dense layers. The embedding layer has 50 units, both LSTM layers have 64 units, and the dense layers have 32 and 1 unit (1 unit for the output layer since this is the final valence). I tried adjusting the activation function for the output layer from sigmoid to softmax, but I found sigmoid worked better for my model. I kept the activation function as relu for the hidden layers.

6. If applicable, indicate the loss and optimizer for your machine learning model.

I used binary cross entropy for the loss and RMSprop for the optimizer.

7. How did you split the training, validation (if applicable), and test data?

I did a 80/20 split between training and testing data. I searched online for typical split distributions, and I found that this was the most common ratio. I also reduced the tweets to one-tenth (160,000 tweets) since training the model took too much time with 1.6 million tweets.

8. What evaluation metrics did you use to evaluate your model? What were your results?

I used binary accuracy to evaluate the model, and I obtained a 76% accuracy.