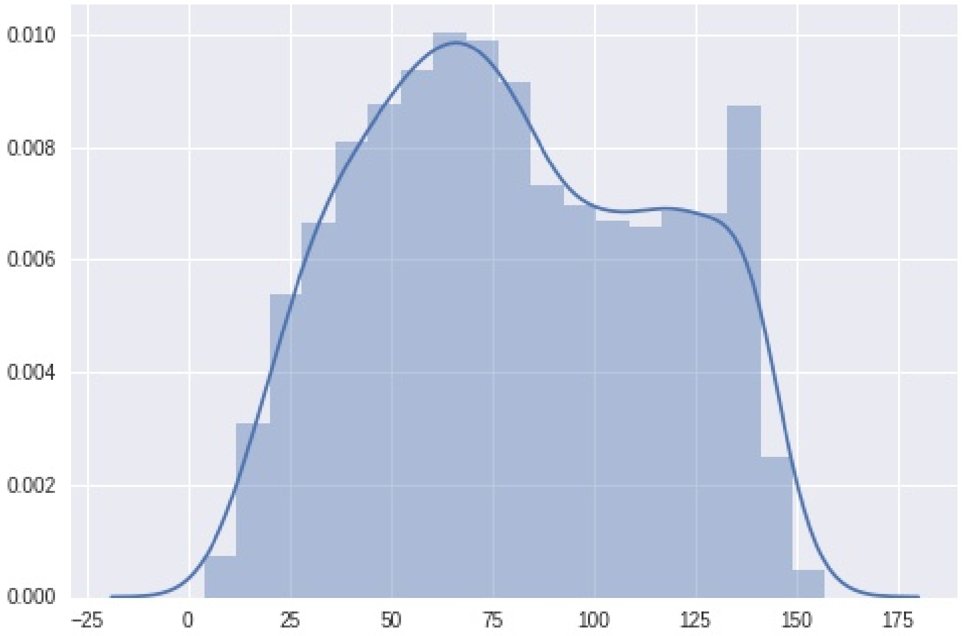
Data Collection:

- Collected the tweets dataset from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/download/13118/12795
    - Could collect only 3800 tweets out of 5800 tweets in the dataset. Rest of them seem to be deleted by the users.
    - Contains three types of questions, i.e.
        - Q1: Does the tweet contain any mention of alcohol?
        - Q2: Is the user speaking about himself/herself being drunk in the tweets?
        - Q3: Is the user drunk at the time of tweeting?
    - Q2 and Q3 are conditional on their previous questions, and hence have smaller datasets.
- Requested for the dataset from the other papers (IIT Bombay) but haven't received any reply yet.

Data Cleaning:

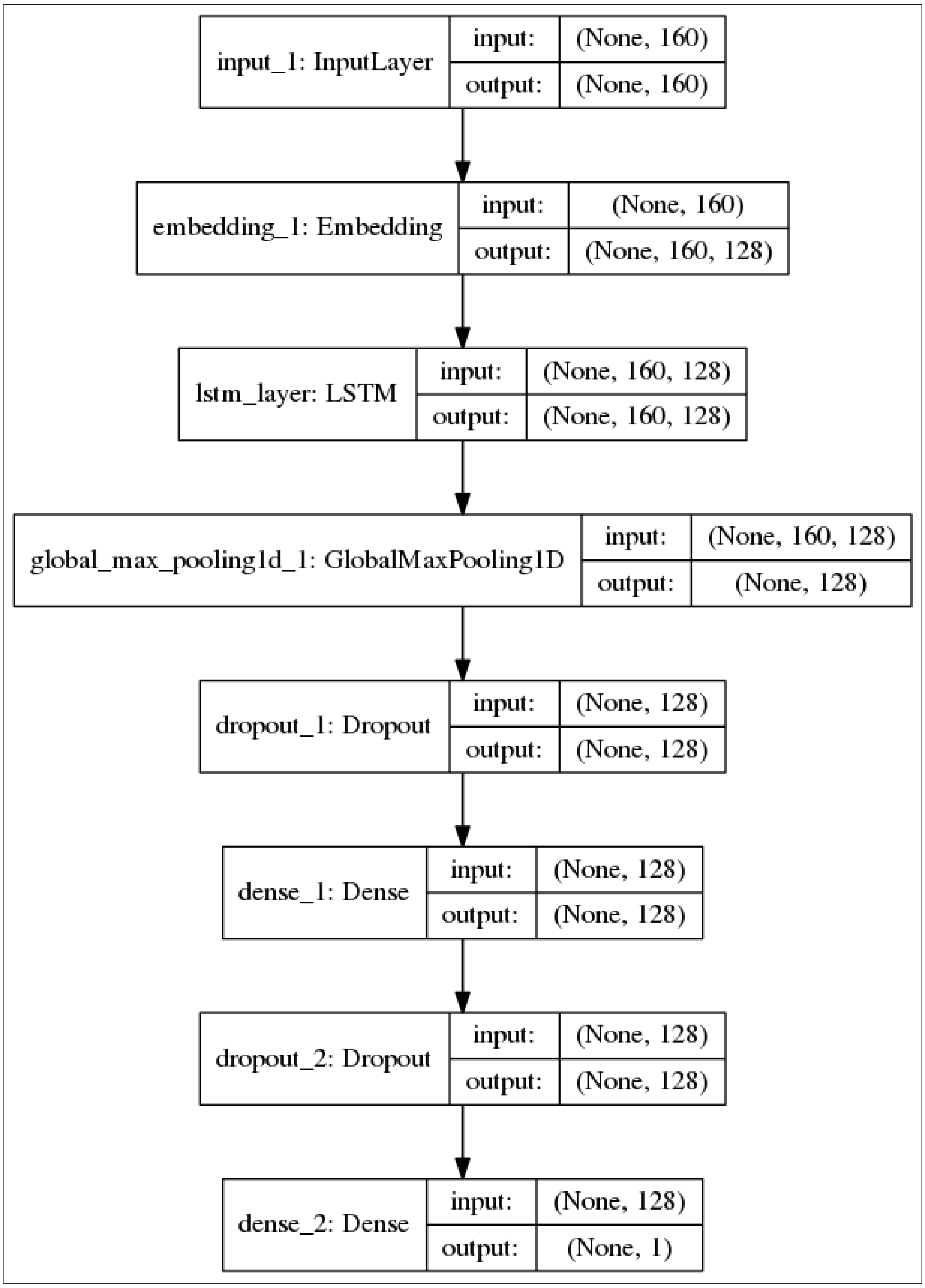
- Did not perfrom any cleaning of the dataset except converting tweets to lower case. The tweets contained emoticons as well which we thought could be important when looking at the sequence of characters.
- Split the dataset into Training and Testing (80-20 Hold-out Validation).
- Tweet length varied from 0-157 characters. The distribution can be seen below:



- Used sequence of characters as input to the model. Encoded every character into an integer including spaces, punctuations and emoticons. There were a total of 133 unique characters.
- Used maximum padding of 160 characters to create the input to the model.

Modeling:

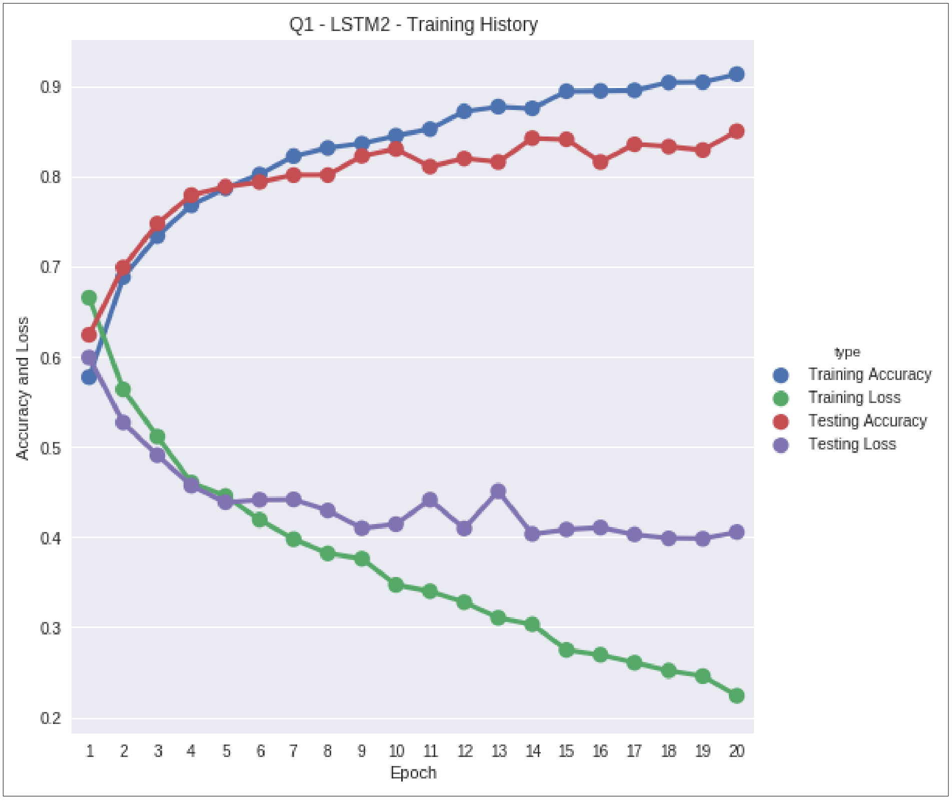
- Architecture of the LSTM Model:



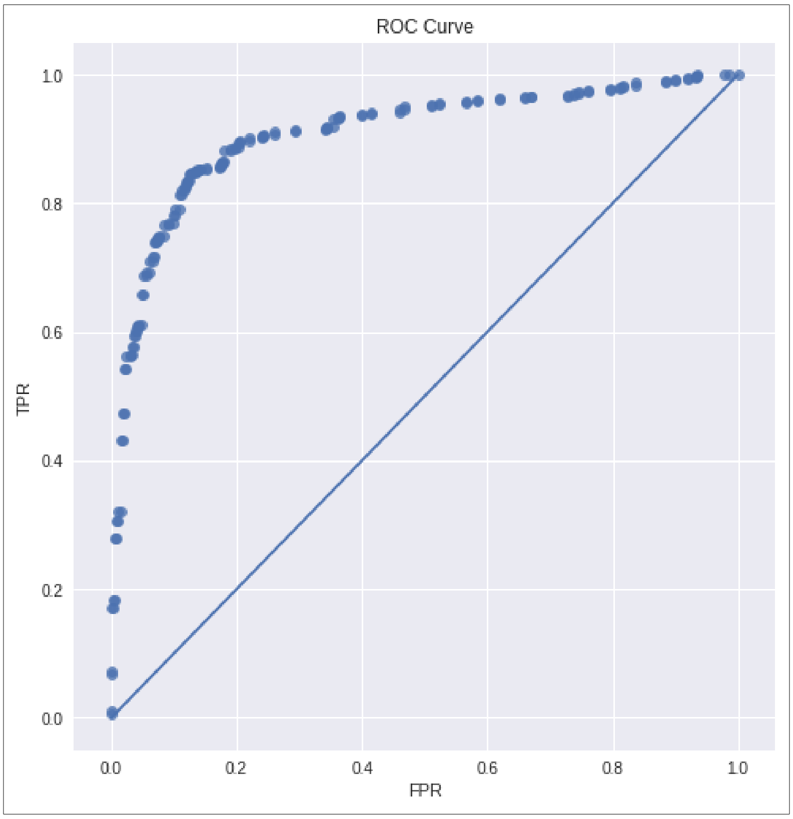
- Using embedding in the first layer improves the accuracy of the model from 62% to 85%
- Other Model hyperparameters:
  - Optimiser: Adam
  - Learning Rate: 0.001
  - Epochs: 20
  - Loss Function: Binary Crossentropy
  - Random shuffling in each epoch
- **Possible Changes:**
  - Data Cleaning
  - Sequence of words instead of characters
  - Tuning hyper-parameters using Cross-validation
  - Create text mining features mentioned in both the papers

Results:

- Training History:



- Loss on testing dataset stabilises after 15 epochs.
- **Classification Metrics:**
  - **AUC: 0.91**
  - **Accuracy: 0.85**
  - **Precision: 0.90**
  - **Recall: 0.82**
  - **F-1 Score (Micro): 0.85**
- ROC Curve:


ROC Curve
TPR
FPR