Unit 4 Evaluation - Group N3

Objective: To build an analyzer to process the performance of the banks with respect to recent government decisions.

Inputs:

1. First set of inputs had seven files, where we separated the entire dataset according to banks. In this dataset, we found 7 banks (HDFC, ICICI, SBI, Kotak, Axis, RBI, Bank Of India). Each of the 7 banks had a file with corresponding tweets along with the label of the tweet(displeasure or compliment or miscellaneous). The banks were named as @bank_name, so tweets which had that name were mapped to that particular bank.

2. Second set of input had three files, where we separated tweets according to the class labels (displeasure, compliment and misc).

Design:

1. Design1 : Using RNN

 - Each bank data is trained and tested using vanilla RNN.
 - That means that, each bank has it's own model, trained, validated and tested by vanilla RNN.
 - In this case, we will have seven RNN models, one for each bank. Basically, this means that we would have trained seven sets of weight matrices and bias vectors.
 - The input given to RNN is a sentence (arbitrary length with words represented as vectors) and output to the RNN is a vector with size 3 which represents whether the tweet was a compliment, displeasure or misc.
- Each RNN is trained first and later tested for accuracy. For testing purpose, we pass test data pertaining to that particular bank only. Using this we can calculate the accuracy.
- For ranking the bank, while testing we count the number of tweets predicted as compliment and number of tweets predicted as displeasured, calculate the compliment/displeasured ratio. The testing is done for all seven models. The model with highest compliment:displeasure ratio is the highest ranked bank.

2. Design2 : Using MaxEnt

  - Similar to RNN, each bank data is trained and tested on it's own model.
  - We have to train the vector dot product, ie, (v . f(x,y)), where v is the model parameter to be trained and f(x,y) is the feature function where, x is the input (say sentence) and y is the target class. Feature functions could be those words which occur many times pertains to that class. For example, if Thank you occurs maximum times in the tweets pertaining to compliment, presence of that word can be an indication that the tweet is a compliment.
  - Such presence of words can be used to build feature functions for the three classes as mentioned.
  - Once feature functions are written, we train the model for seven different banks using maximum likelihood estimation.

Log Linear Model

$$P(y|x) = \exp(v.\ f(x,y))/\Sigma \exp(v.\ f(x,y))$$

- Once, the model is trained, the model is tested for accuracy. The number of compliments, displeasures and misc were found and the banks were ranked based on compliment/displeasures value.

Procedure:

1. Preprocessing:

As mentioned, the input files for banks and classes were created after cleaning the tweets. Cleaning tweets involved removing hashes and other random characters, analyzing the document to write feature vectors for maxent, extracting bank names (using regex).

2. Model:
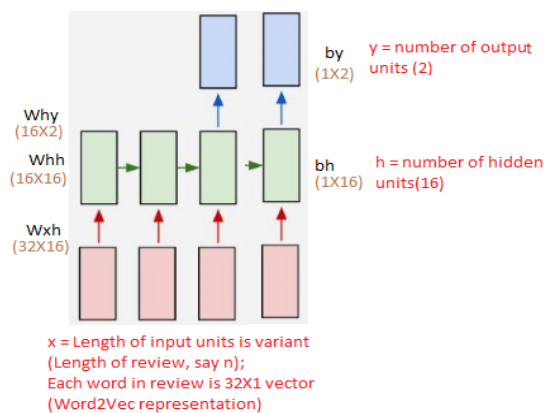
RNN

In the figure above,

black - The Parameters of the model used
brown - The dimensions
red - The model

Hyperparameters:

1) Number of units in each layer:
       - Input: Length of sentence; variant;
       - Word representation: 32X1 vector
       - Hidden : 16 units
       - Output : 3 units (First unit represents negative and second unit represents positive)
2) Learning rate: 1e-1
3) Epochs: We tried with 100, 500 and 1000

2. Output:
  - Preprocessing took time till 6.30PM as it was throwing exception because of missing tag in the dataset. Took time to realize where the exception was from.
  - With RNN the current accuracy is 70.4%.

Improvising:

1. Yet to try MaxEnt.
  The results given by MaxEnt depends on the feature functions written and the relevance of test data with respect to trained data.

2. For hybrid model, we can try passing output of MaxEnt, that is, the predicted class as input to RNN. RNN then predicts the bank from which the tweet was in reference to.

For example, if there are 20 tweets predicted as compliment by MaxEnt and 50 tweets predicted as displeasure by MaxEnt.
Once we pass the 20 tweets (compliment) as input to RNN, the RNN predicts which bank the tweet corresponds to. So, out of 7 banks, 12 tweets predict B1, 2 tweets predict B2, ... k tweets predict B7 (12+2+..+k=20).

Similarly, test the same way for displeasures.

Once we get, compliment and displeasures, calculate compliment/displeasures ratio and order the banks accordingly.