SENTIMENT ANALYSIS

CA 2 REPORT

PROJECT	SUBMITTED BY	
ROLL NO	NAMES	REG NO
47	DEVANSH CHITRANSH	11812006
45	MERUGU RANJITH	11811946

SECTION :KM109

SUBJECT: SOFT COMPUTING TECHNIQUES

ROLL NUMBERS: 47 , 45

LOVELY
PROFESSIONAL
UNIVERSITY

COMPUTER SCIENCE ,ENGINEERING

LOVELY PROFESSIONAL

UNIVERSITY,JALANDHAR

OCTOBER -2020

SECTION : KM109

STUDENT DECLARATION

This is to declare that this report has been written by us.No part of the report is copied from other sources.All information included from other sources have been duly acknowledged.No part of the report is copied

We are shall take responsibility for it.

SIGNATURE OF STUDENTS

Devansh chitransh

Merugu Ranjith

Roll no: 47 , 45

Place: Lovely professional university

Date: Oct -10 - 2020

TABLE OF CONTENTS

ABSTRACT

The project that we choosed is" SENTIMENT ANALYSIS " .This project basic motive is to predict the emotion of the text that a user gives .We can say this project helps us to categorize the data set that user gives into different emotions .

INTRODUCTION

In todays era of social media , there are may ways of expressing the thoughts of a person . Suppose a person posted something unusual thing in the social media lets say twitter , There might be different comments to the post . It can be positive , negative or there might be something neutral .

So inorder to categorise the comments of some post into different emotions such as positive comments , negative comments , neutral .
So to do this thing we have created this project .

To do this project we used Reccurent neural network . RNN is so useful to recognize the pattern across time , This thing makes our model so accurate to predict the sentiment .

And we used LSTM to take in input size a hidden and no of layers as well this helps to find better probability of finding accurate answer.

Padding sequence also used to deal with short and long reviews .

We do tokenization of text after applying required algorithm to make it simpler so the training model predict the near/ almost accurate answer of the reivew /comment given by user .

LITERATURE REVIEW:

RECCURENT NEURAL NETWORK:

It is a class of artificial neural network where connections between nodes form a directed graph along a temporal sequence.This allows it to exhibit temporal

dynamic behaviour .Derived from feedforward neural networks,RNN can use their internal state (memory) to process variable length sequences of inputs .

LONG SHORT-TERM MEMORY(LSTM):

It is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video).

BATCHING:

The batch size is a hyperparameter that defines the number of samples to work through before updating the internal model parameters.

A training dataset can be divided into one or more batches.

Batch Gradient Descent. Batch Size = Size of Training Set
Stochastic Gradient Descent. Batch Size = 1
Mini-Batch Gradient Descent. 1 < Batch Size < Size of Training Set

PROPOSED METHODOLOGY (ALGORITHM)

Why RNN ? And why not CNN?

An RNN is trained to recognize patterns across time, while a CNN learns to recognize patterns across space. Which DNN type performs better when dealing with text data depends on how often the comprehension of global/long-range semantics is required.

Step 1:

we used data of us airlines twitter review data from kaggle website.

Step 2:

Load the data and visualize it.Just check the overall review column and row wise.

Step 3:

Data preprocessing initially we are using embedded layers so lets make it into proper form to feed the network .
Encode each and every layer as we using emebedded layers. split all text into words.

Step 4:

Encode the words according to the words that got splitted create a dictionaires of map the vocabulary to integers .then convert each integer into review.

Step 5:

Encode the labels identify tweet as negative or non negative if its positive label it as 1 otherwise 0 accordingly.

Step 6:

Padding sequences if the review is so long do truncate of that otherwise padd the review , We can say that a good length of tweet is 30 character max review length is 32 as we dealing on lesser size of texts.

Step 7:

Now almost half the part of tokenization and all completed now we use RNN network to train the data set that we modified .

Step 8:

After creating training ,test and validation data we can create data loaders for this data by following two steps:

a) Create a known format for processing data ,using tensor dataset which takes in a input set of data and a target set of data with the same first dimension ,create a dataset

b) Create data loaders and batch our training validation ,and test tensor data sets;

Step 9: Sentiment network with pyTorch

We used embedding layer and do add lstm to RNN to take input and analyise it .

Step 10 :

We just check wether RNN is working or not if yes we proceed further.

Step 11:

we will use a cross entropy loss which is designed to work with single sigmoid output .

Step 12 :

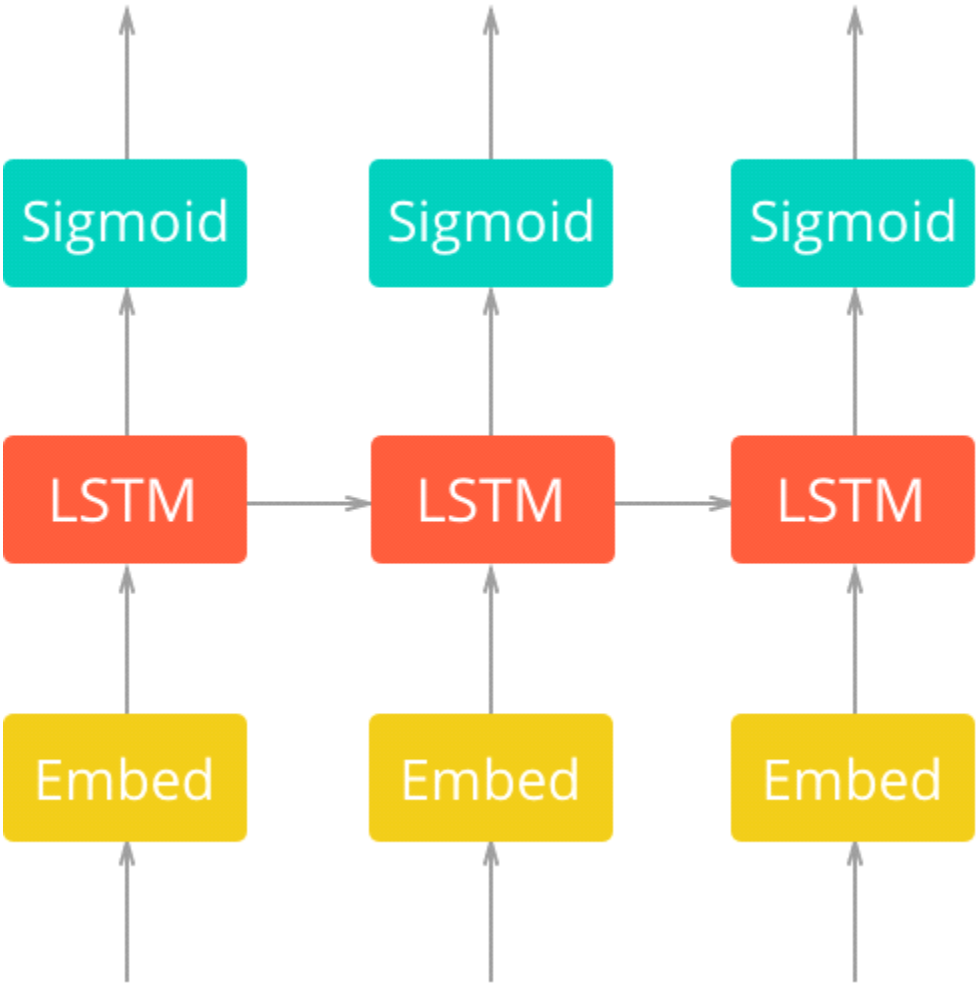
Do test the accuracy of output and average loss of data .(No model can assure 100 per accuracy and 0% loss)

Step 13 : Final step

Create a predict function which can predict the data accurately . And gives us the output either postive, negative or neutral .

NEURAL NETWORK DIAGRAM :

Following figure resembles the RNN



RESULTS AND DISCUSSION

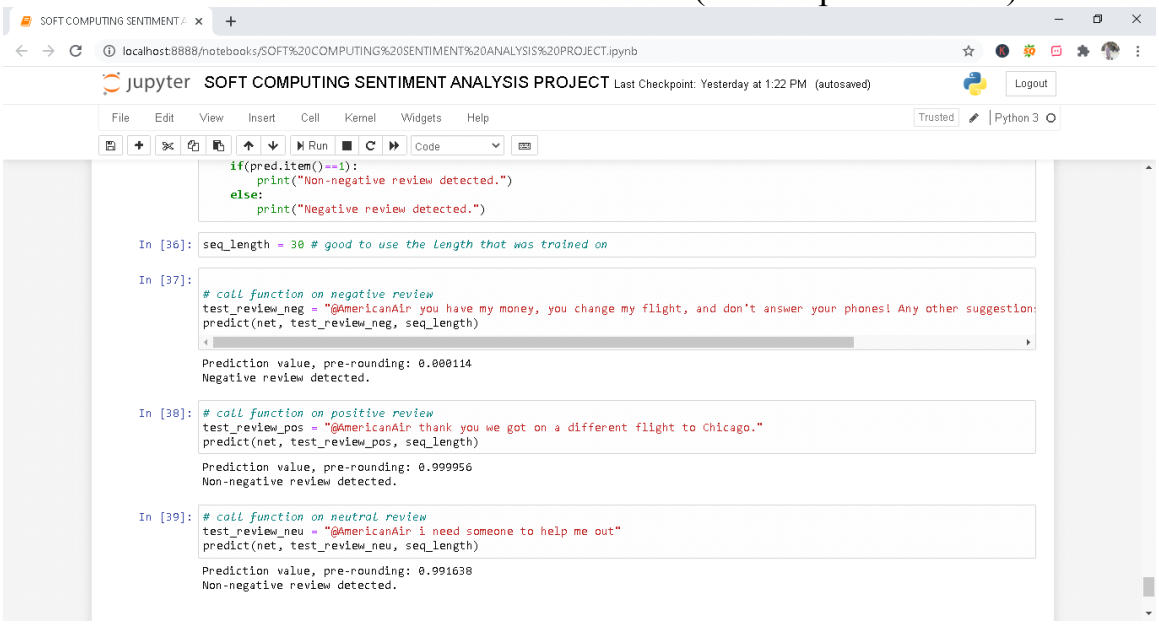
INSTRUCTIONS TO EXECUTE THE PROJECT

There are several modules which should be compiled before moving to the final input ouput program.
First check wether data loaded or not.
Then check the wether data got pre processed or not.
Then check the output of padding sequences and check wether output is in integer values or not.
Then check the trained data.

Then check the working of RNN .
Finally do validata the tested , encoded data.
Finally give user input in predict function and check wether its positive or not.

SAMPLE INPUT AND OUTPUT:

THIS IS THE JUST A INPUT OUTPUT PIC (1/10 th part of code)



FOR OVERALL CODE AND INPUT OUTPUT CHECK THIS LINK:
https://drive.google.com/file/d/1XVG2i5koAebGZwrdgqZjHKB0Y27UxiLx/view?usp=sharing

CONCLUSION:

From this project we can conclude these

This project is so much helpful for grouping reviews of airlines into different categories .

The positive review help them to maintain the things in good way .

The negative review help them to improve their things .

Reccurrent neural network makes this project much more accurate than it was .

The batching padding is so much useful in splitting comments that are either large or short .

This project helped us to understand the difference between RNN and CNN .

And working of predict function as well as working and usage of pytorch .

This project not only made us to build sentiment analyser it also helped us to learn concepts in better and practical way .