****

**Ahsanullah University of Science and Technology**

**Department of Computer Science and Engineering**

**Course No. :** CSE 4238

**Course Name :** Soft Computing Lab

**Assignment No. :** 03

**Submitted By:**

Name : Emdadul Haque

ID No. : 17 01 04 028

Session : Fall - 2020

Section    :       A (A2)

Table of Contents

[1. Dataset Creation 2](#_Toc83481229)

[1.1. Data Downloading: 2](#_Toc83481230)

[1.2. Creating Own Dataset 3](#_Toc83481231)

[1.2.1. Read the raw data: 3](#_Toc83481232)

[1.2.2. Dataset splitting 3](#_Toc83481233)

[1.2.3. Text cleaning 4](#_Toc83481234)

[1.2.4. Saving the new data 4](#_Toc83481235)

[2. Training Process 5](#_Toc83481236)

[2.1. Load Dataset 5](#_Toc83481237)

[2.2. Model Creation 6](#_Toc83481238)

[3. Result 9](#_Toc83481239)

[3.1.1. Parameter 9](#_Toc83481240)

[3.1.2. Loss graph plotting 9](#_Toc83481241)

[3.1.3. Accuracy graph plotting 9](#_Toc83481242)

[3.1.4. Confusion matrices 10](#_Toc83481243)

[3.1.5. Other’s performance matrices 10](#_Toc83481244)

[6. References 10](#_Toc83481245)

# 1. Dataset Creation

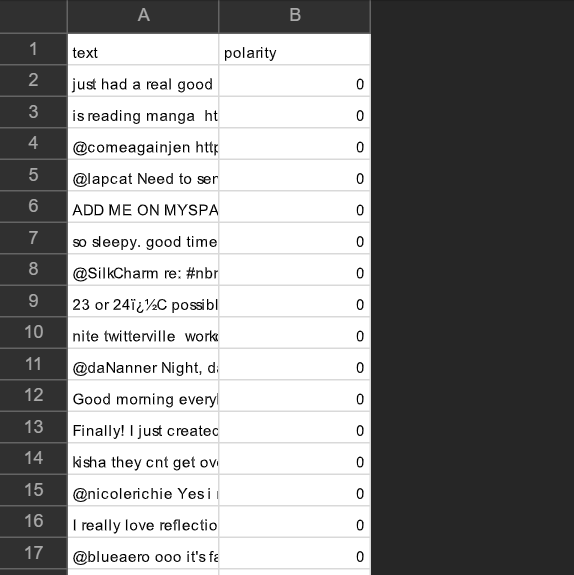
In this section, we know about the creation process of the dataset.

For the data creation process, we divide it into two different sub-processes.

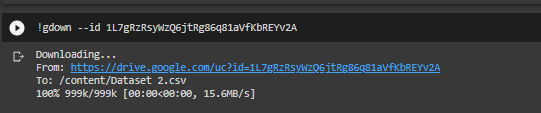
1. Data Download
2. Creating Own dataset
   1. Read the raw data
   2. Dataset splitting
   3. Text cleaning
   4. Saving the new data

## Data Downloading:

In the data downloading process, first, we need to download file based on my id. So, my id is 028. So, 28%3 = 1. So, my dataset number is 2. I need to download [Dataset 2](https://drive.google.com/file/d/1L7gRzRsyWzQ6jtRg86q81aVfKbREYv2A/view). The dataset looks like-



For download the dataset I used gdown package. Because this is google drive link. For downloading I used given code-



## 1.2. Creating Own Dataset

After Collecting the raw data. We need to generate and pre-process dataset.

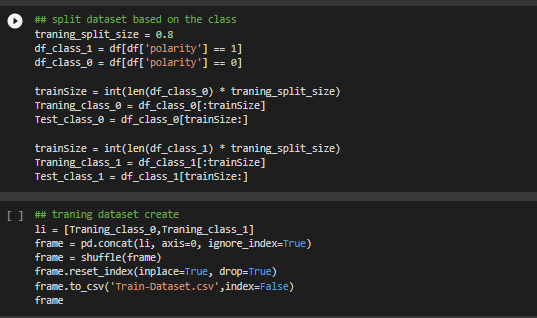
### 1.2.1. Read the raw data:

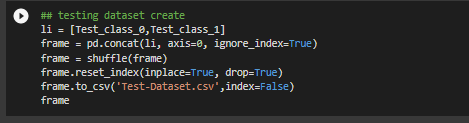
For read the csv file I use pandas dataframe. Because it is easy to use. For read data properly. I also need the encoding. For reading the raw data I use this code –



### 1.2.2. Dataset splitting

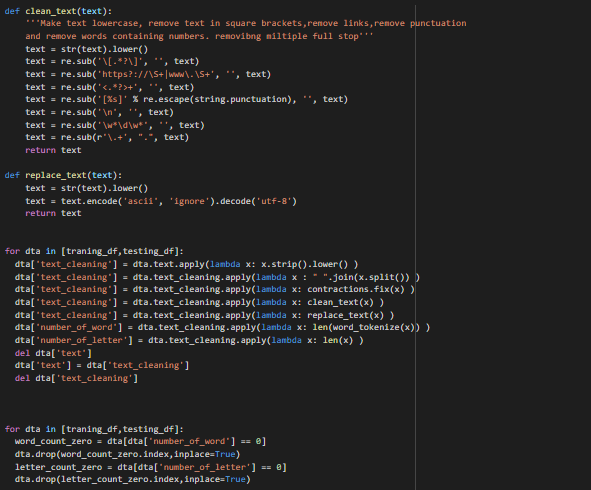
After reading data I need splitting data training and testing. Also, I will maintain 80% and 20% data. Also, I try to maintain positive and negative class properly. The code looks like –





### 1.2.3. Text cleaning

In our dataset we used text-based dataset. We know text preprocessing is an important part of data of NLP. That’s why we need to preprocess dataset properly. For text cleaning we used this code.



### 1.2.4. Saving the new data

After doing those process, we must need to save the data. For saving the data we used this code –



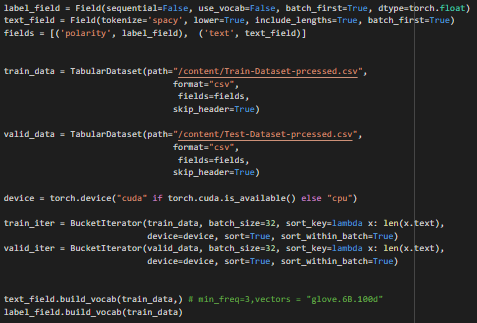
# Training Process

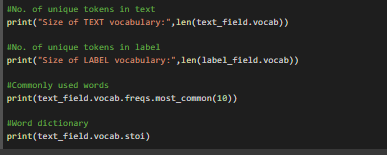
For training process, I follow those steps –

1. Load dataset
2. Model Creation

## Load Dataset

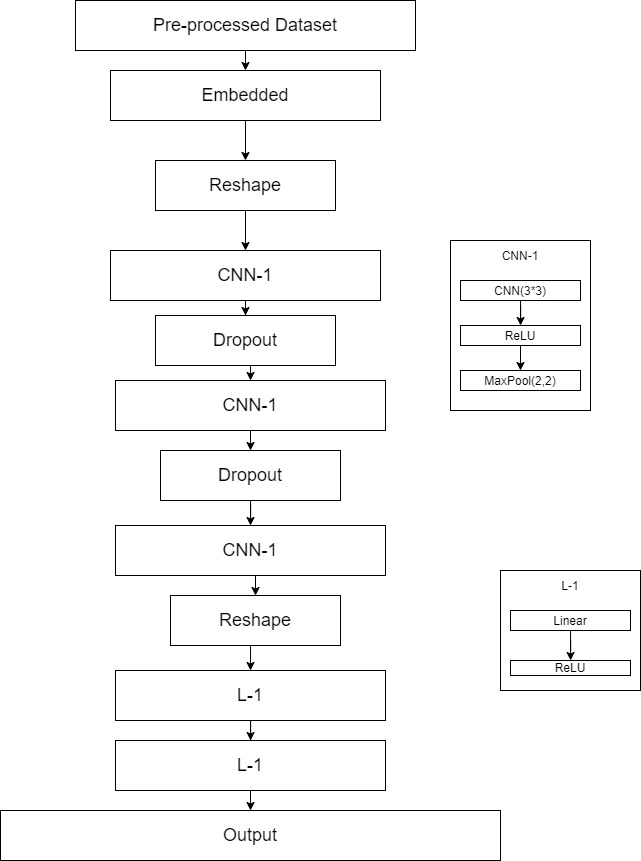
For loading the dataset I used TabularDataset and BucketIterator. TabularDataset is use to read data from saved and pre-processed. The dataset loaded code-



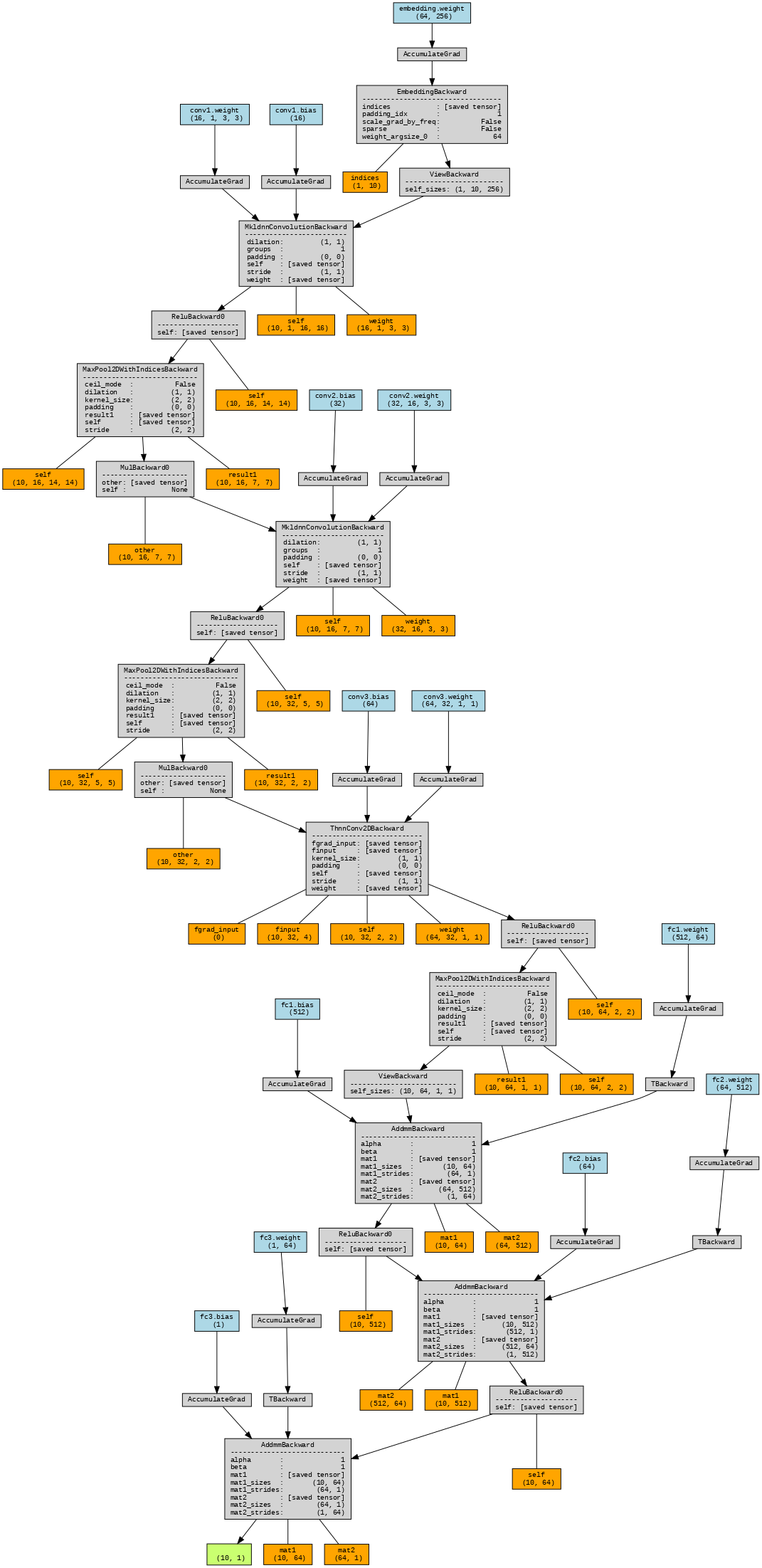


## Model Creation

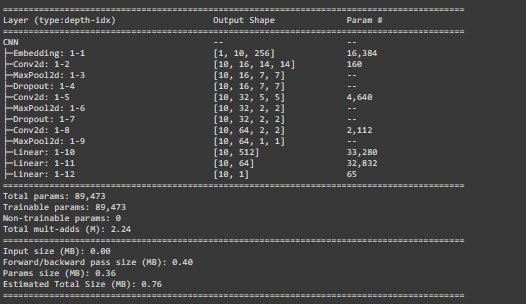
For the question, I need to make the CNN based model. My model architected look like.



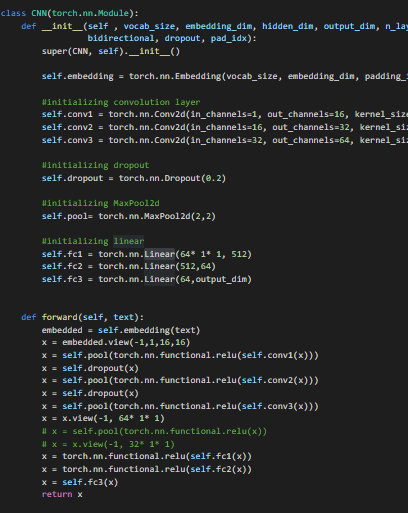
The flow of model is –



The model parameter is –



Also, the code of model is-



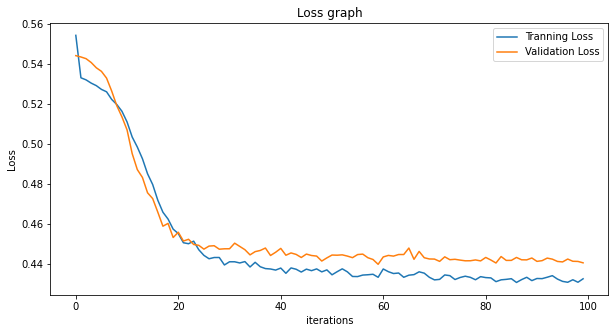
# Result

For training the data I will use Colab. Because my personal computer is not enough to load this high performance. Below I was showing all the training result.

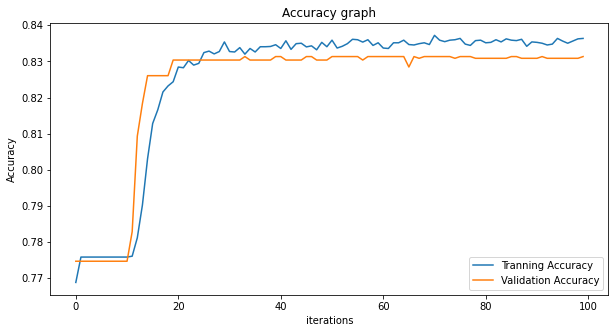
### 3.1.1. Parameter

|  |  |
| --- | --- |
| **Hyper Parameter** | **Value of parameter** |
| K(factor) | 10 |
| iteration | 100 |
| Platform run | Colab |

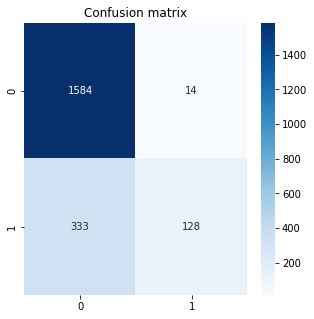
### 3.1.2. Loss graph plotting



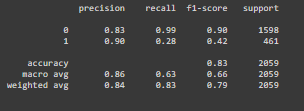
### 3.1.3. Accuracy graph plotting



### 3.1.4. Confusion matrices



### 3.1.5. Other’s performance matrices



# 6. References

Source code:

<https://colab.research.google.com/drive/162ys8eXTEQi6Lj6CH6EKT_C88F9RnYlw>

Data info:

<https://drive.google.com/drive/folders/1-PImfvx9IGyMHo7ryZoFXgCrfRLlKu9K?usp=sharing>