**Deep Learning - Lab ASSIGNMENT 2**

**Name:** Rajeswari Devi Namana

**Class ID:** 29

**Date:** 11/16/2017

**I. Introduction:**

The task given for this lab assignment is to implement text classification with CNN model with new data set (minimum 5 classes) which is not used in class. CNN stands for Convolutional Neural Networks. CNNs are used to classify or analyze either images or text. The detailed implementation of CNN to classify text is given below.

**II. Objectives:**

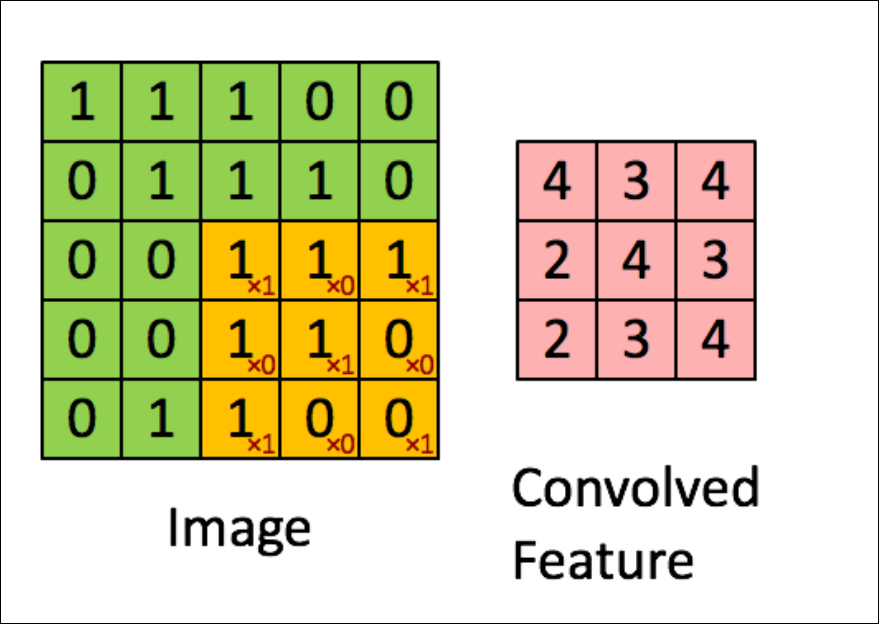
The objective is to train the model to classify text into five classes by using Convolutional Neural Networks. The classes/categories I have chosen to classify text into are: Politics, Entertainment, Sports, Business and Technology.

**III. Approaches/Methods:**

Convolutional Neural Networks approach is adopted. In this model there are three layers which are input, output and hidden layers. Hidden layers are present between input and output layers, and they consist of convolutional layers, pooling layers, fully connected layers and normalization layers.

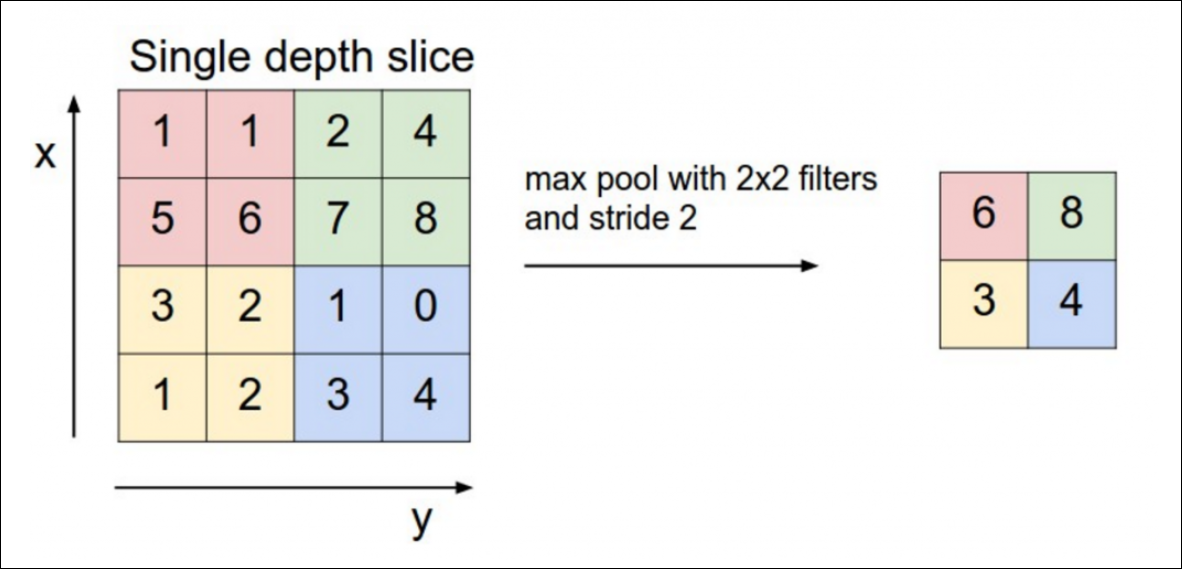
**1. Convolution Layers:**

Convolution can be considered as a sliding window function applied to the input matrix which represents an image or a sentence.



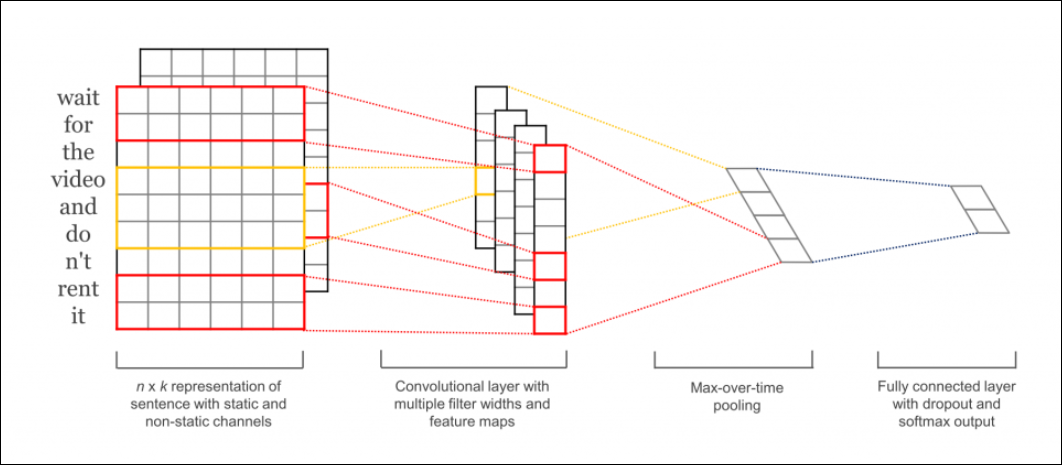
**2. Pooling Layers:**

They are generally used for dimensionality reduction of the matrix and are applied after convolution layer. The most common pooling method is max pooling where a certain stride size is selected and the maximum element for each stride size of the matrix are chosen to produce the output.



**3. Fully connected Layers:**

They are used to connect each region of the input to a neuron in the output. The complete CNN depiction is as follows:

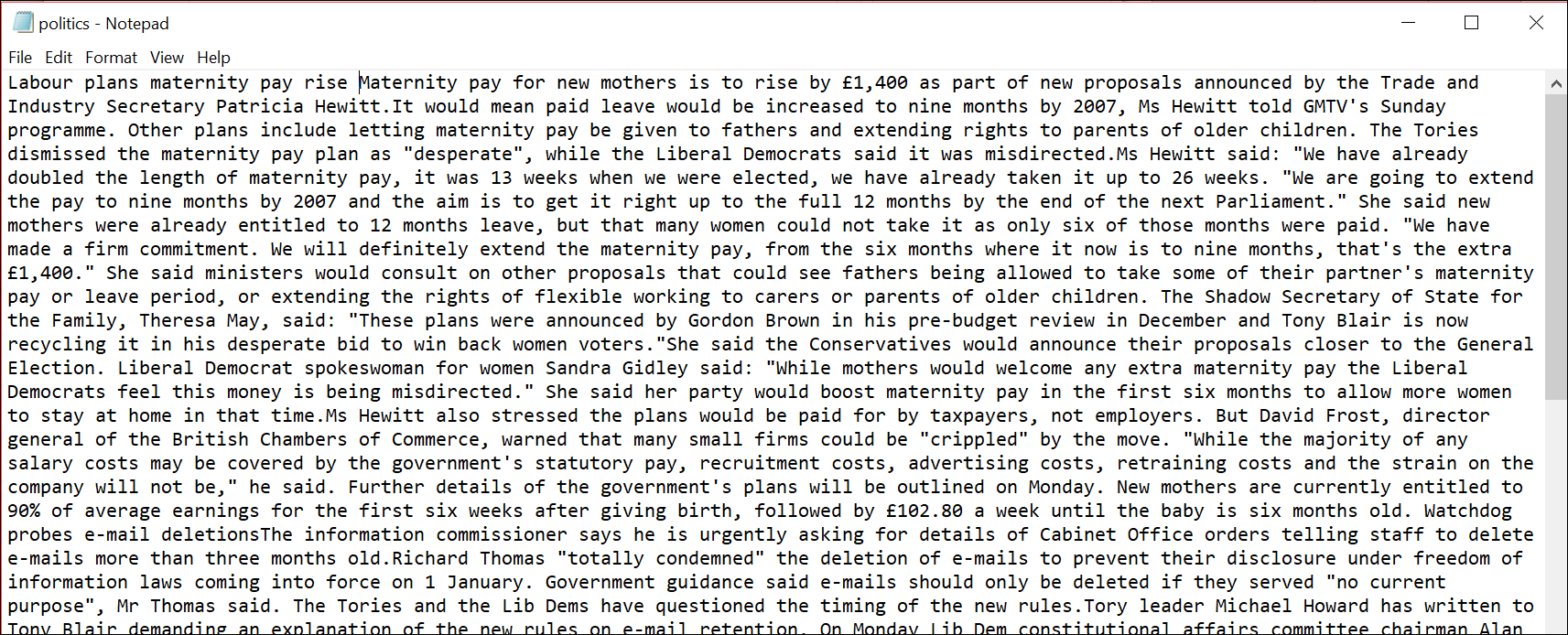


**IV. Datasets:**

I used five different datasets which are gathered from BBC news to train and test the model. The information has been taken from <http://mlg.ucd.ie/datasets/bbc.html> . The 5 categories of data are related to Politics, Entertainment, Sports, Business and Technology. A snapshot of the datasets are as follows:

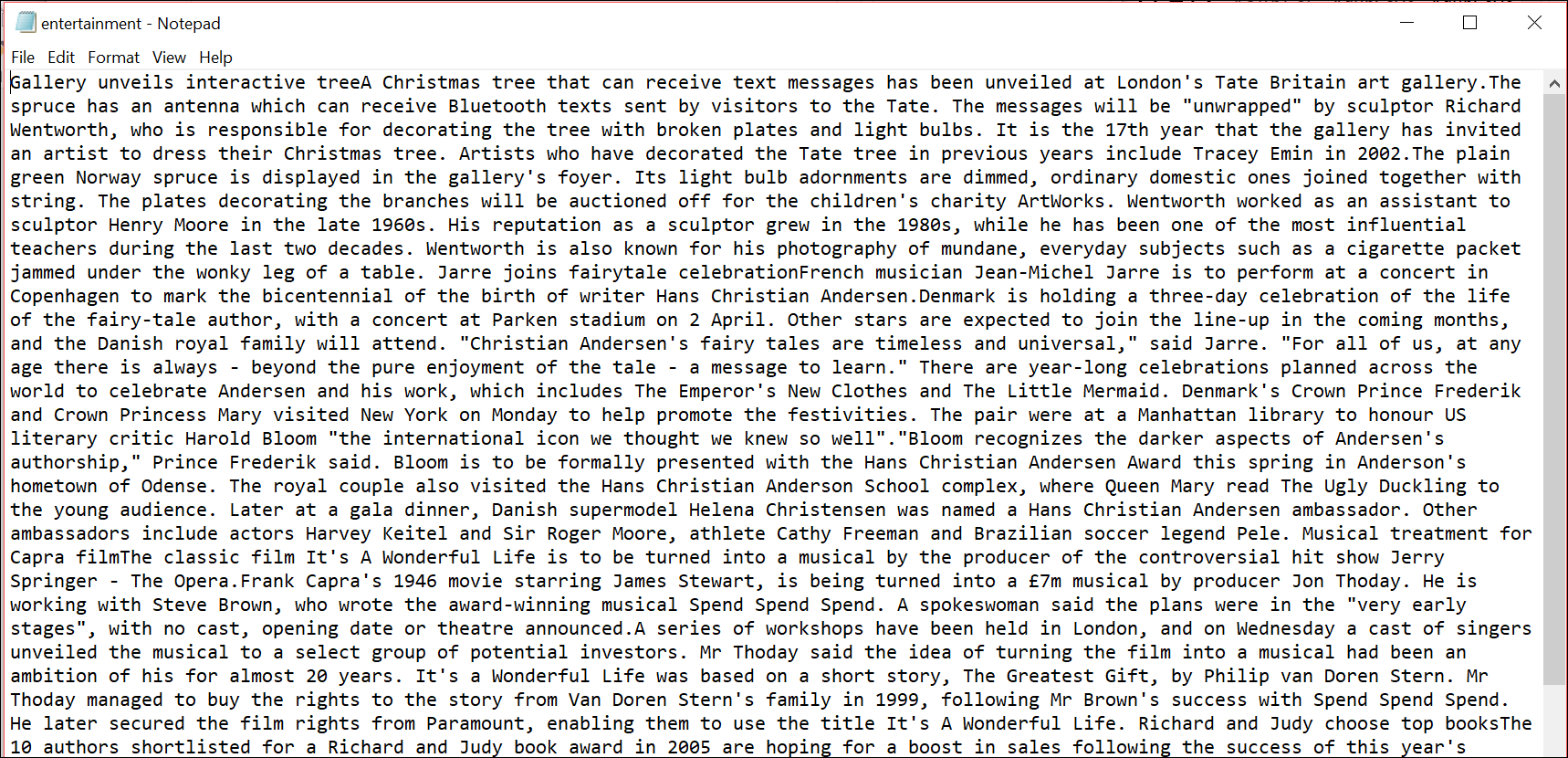
1. **Politics Data set:**

This dataset is about political issues, politicians, their policies, etc.



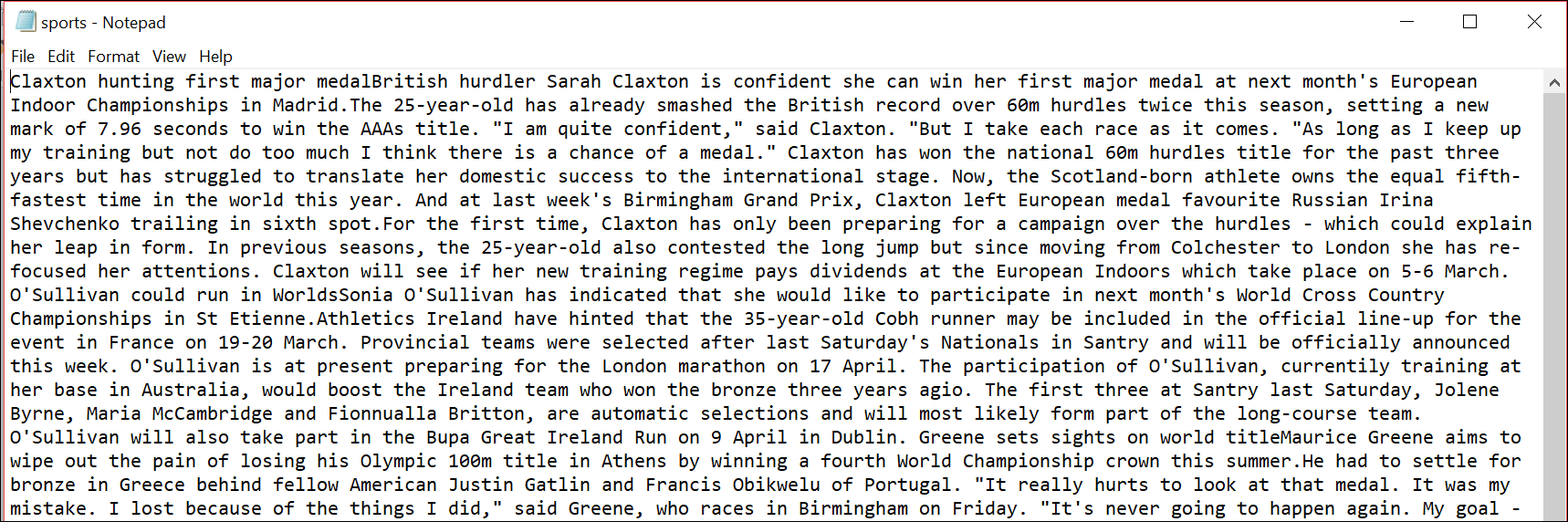
1. **Entertainment Data set:**

This dataset is about various entertainment stuff, stuff related to artists, sculptors, movies etc.



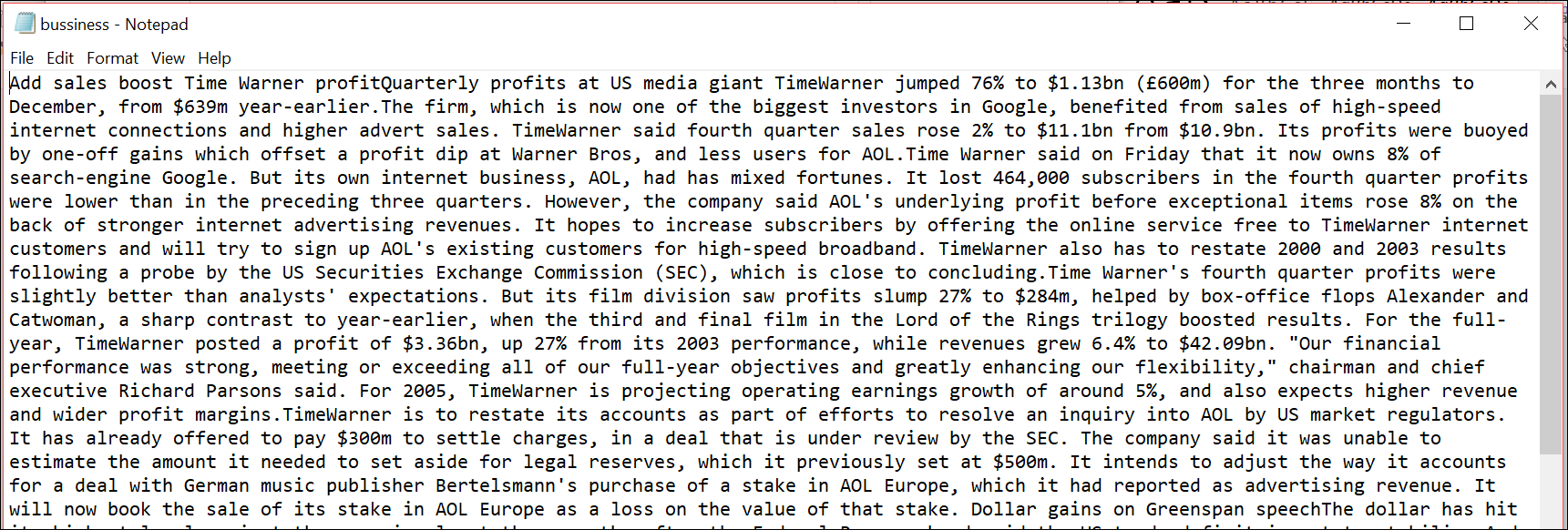
1. **Sports Data set:**

This data set consists of different sports news related to European Indoor Championships, Olympics, Norwich Union Grand Prix, etc.



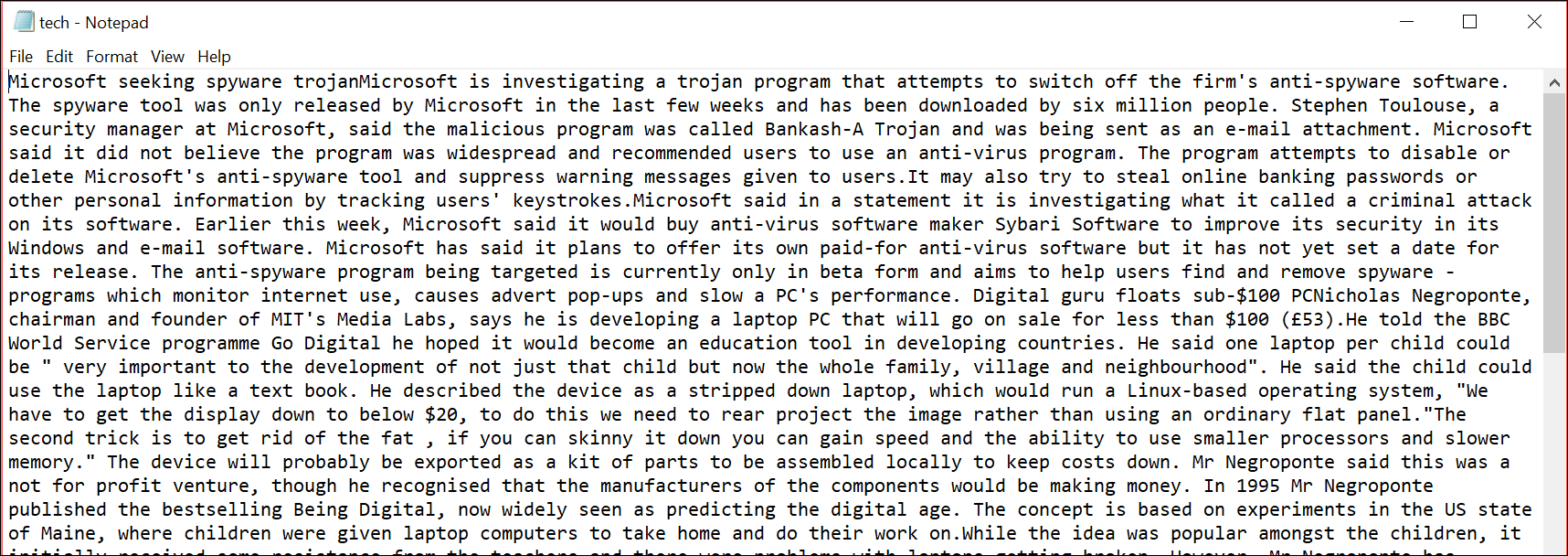
1. **Business Data set:**

This dataset consists of business news related to different companies, their investments and profits.



1. **Technology Data set:**

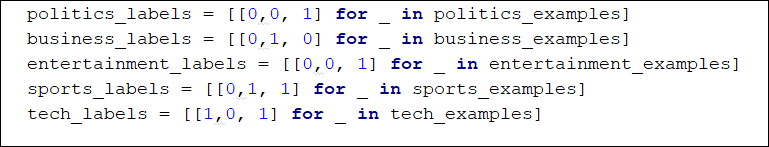
This dataset is about technical stuff like about Microsoft’s spyware tool, $100 laptops which can be used as educational tools etc.



**V. Workflow:**

The implemented CNN model workflow is as follows:

1. First the five different datasets are loaded, split by words and cleaned.
2. They have been labeled into five categories as follows:



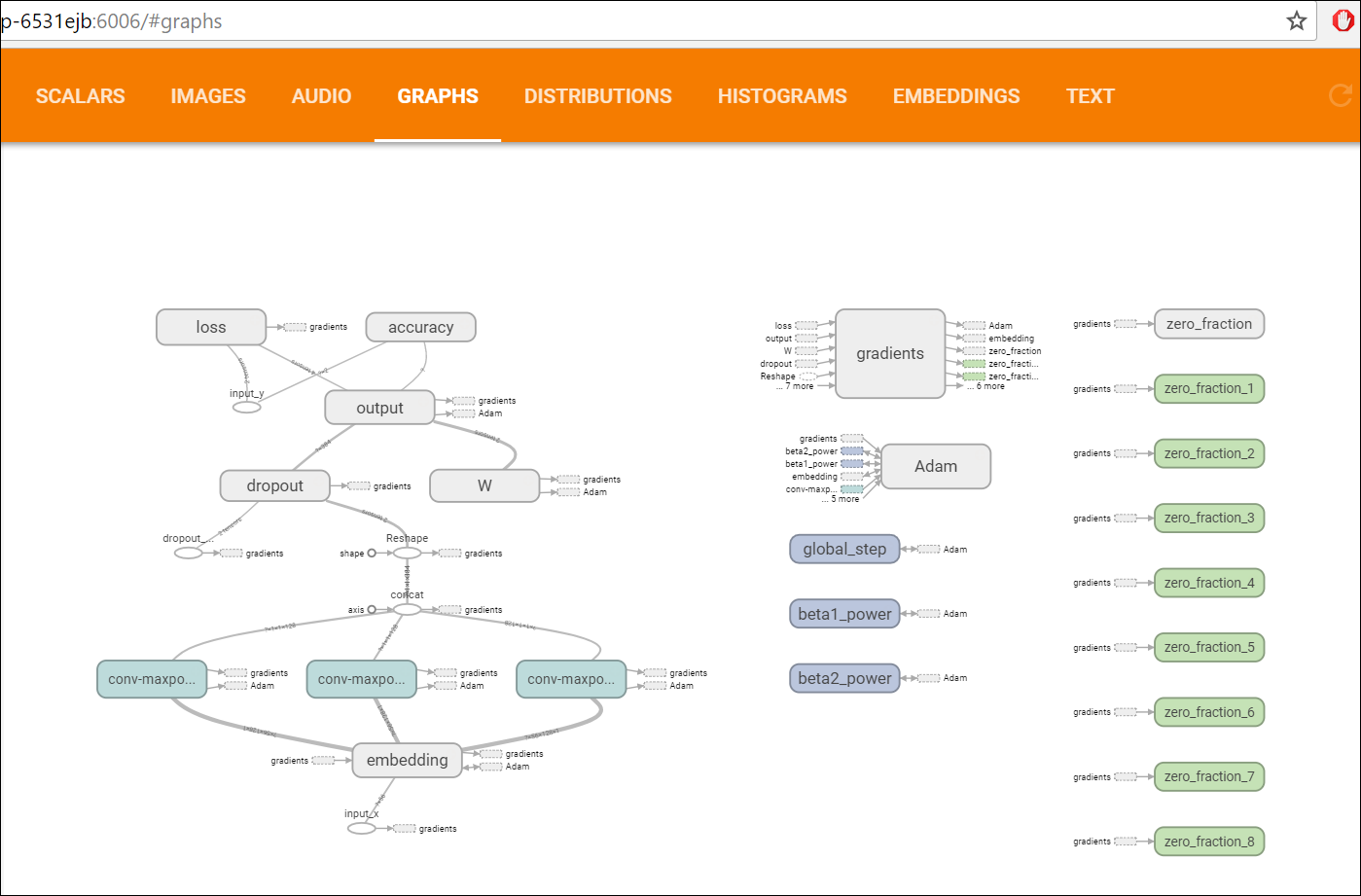
1. All the parameters have been declared to use in the model.
2. Embedding layer has been applied to pad the sentences,
3. Convolution and max pooling are performed which are discussed above.
4. Then dropout layer which disables a fraction of its neurons is implemented. This prevent neurons from co-adapting and forces them to learn individually useful features.
5. After that predictions are made by matrix multiplication and the class with the highest score are picked.
6. Then loses are calculated using softmax cross entropy function and accuracies are also calculated.
7. Finally, the graph is visualized using tensor board.

**VI. Parameters:** The different parameters used are as follows:

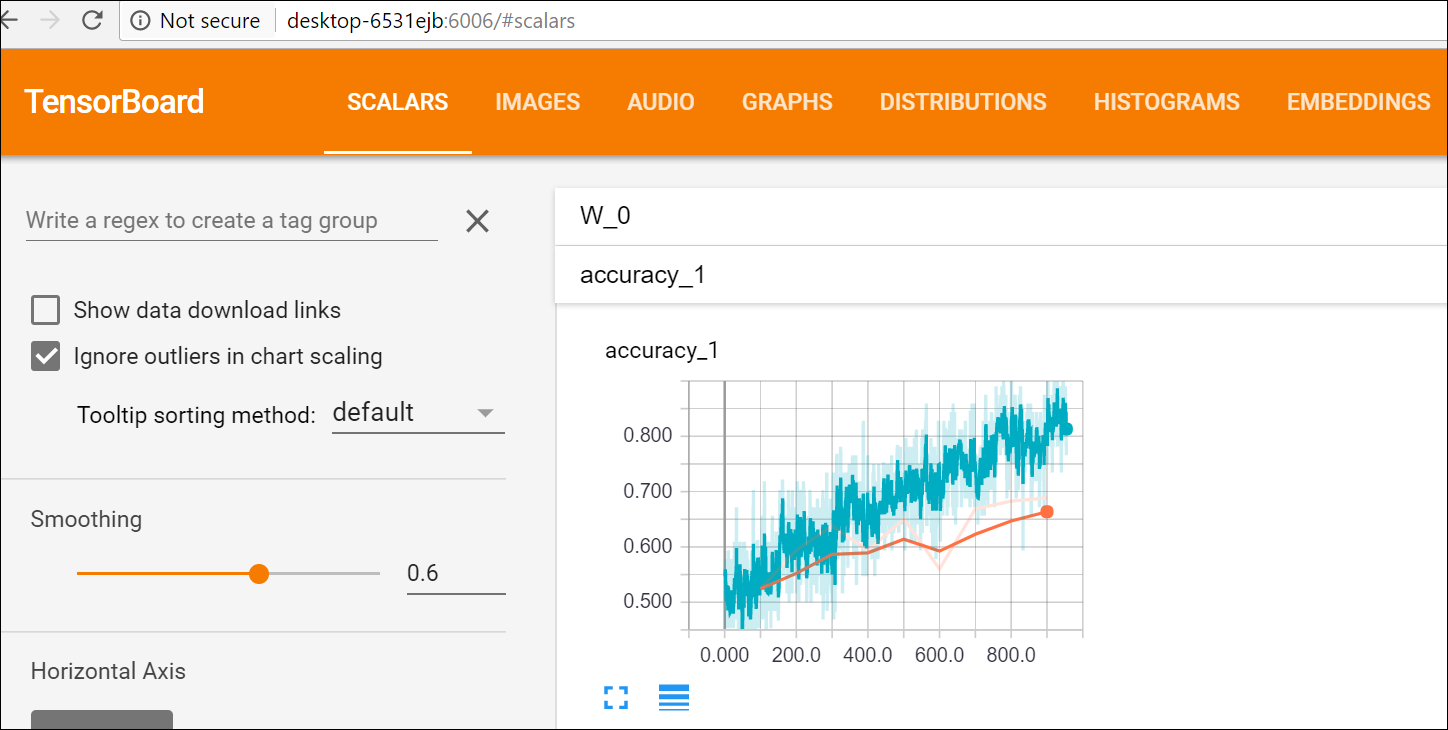
1. Learning rate: How quickly the cost function is adjusted.
2. Training epochs: How many training cycles we go through
3. Batch size: Sizes of training data
4. Sequence length: Length of each sentence.
5. Number of classes: Categories into which text must be classified.
6. Embedding dim: The dimensionality of embeddings.
7. Filter size: The size of convolution matrix.
8. Num Filter: Number of filters used.
9. Checkpoint every: Save model after these many steps
10. Evaluate every: Evaluate model after these many steps

**VII. Evaluation & Discussion:**

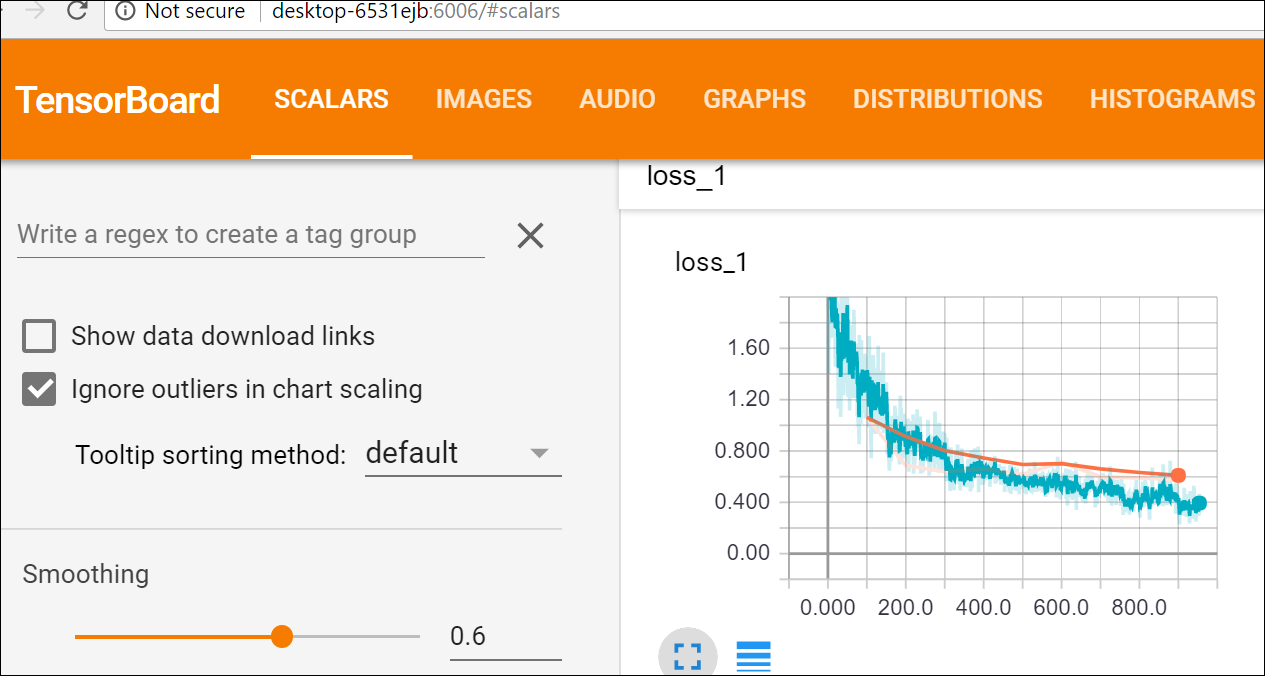
The graph of the model is as below:



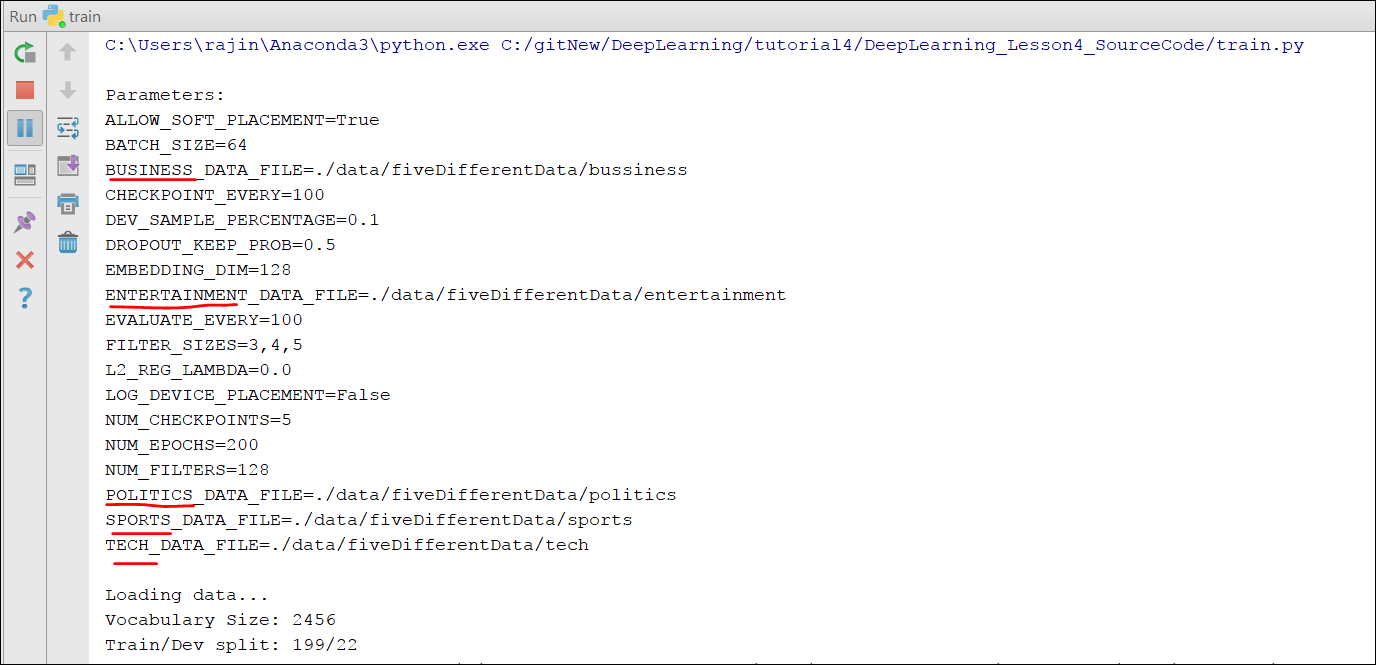
Accuracy:

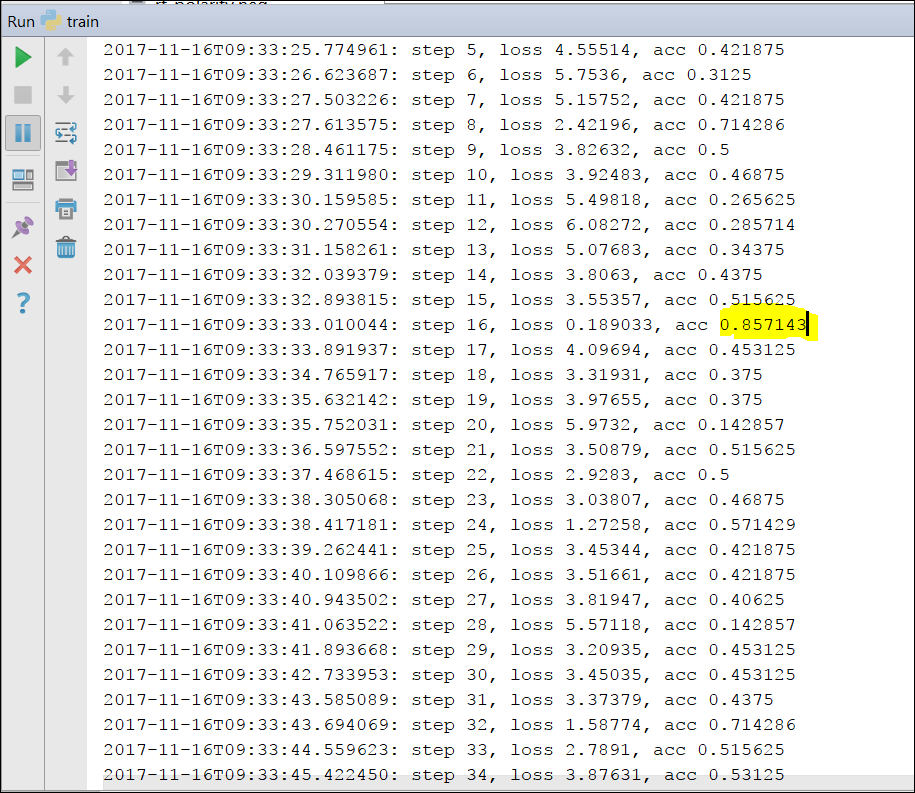


Loss:



Output:





From the output the highest accuracy received is 85%. But there is a lot of variation is the accuracies obtained for each epoch. From the graphs above we can say that since we didn’t achieve good accuracy for testing data when compared to training data, we can say that our model is overfitting the data since the size of dataset is small.

**VIII. Conclusion:**

Convolutional Neural Networks model has been successfully implemented to train and test classification of text into five categories which are Politics, Entertainment, Sports, Business and Technology.