Theme classification for texts

Stanford | ENGINEERING
Computer Science

Albin Forsberg, Rémy Zawislak {albinfor, remzawi}@stanford.edu

Motivation

The motivation behind this project is to find a way of classifying and summarizing the theme of a text. We do so by building a model that learns from labelled quotes to detect the theme of a sentence or group of sentences. The trained model is then used on all the sentences of the text to find the global theme of the text.

Data

The original dataset that is used is the one provided in [1].

- 76 000 quotes
- Author
- 117 labels

A modified version of this dataset was used containing the following information instead.

- 37413 quotes
- 16 labels

For the DNN and NB the words were stemmed, and for the CNN, we used word embedding with pretrained features found in [2].
The data was split into a 80/10/10 train/dev/test set for model evaluation purposes.

Features

- NB and DNN
  - Number of words in a dictionary above a certain threshold
- CNN
  - Sequence of indices of words in a dictionary padded to be fo the same size

Models

The Naive bayes model used is of the standard model with a term alpha added for smoothing.

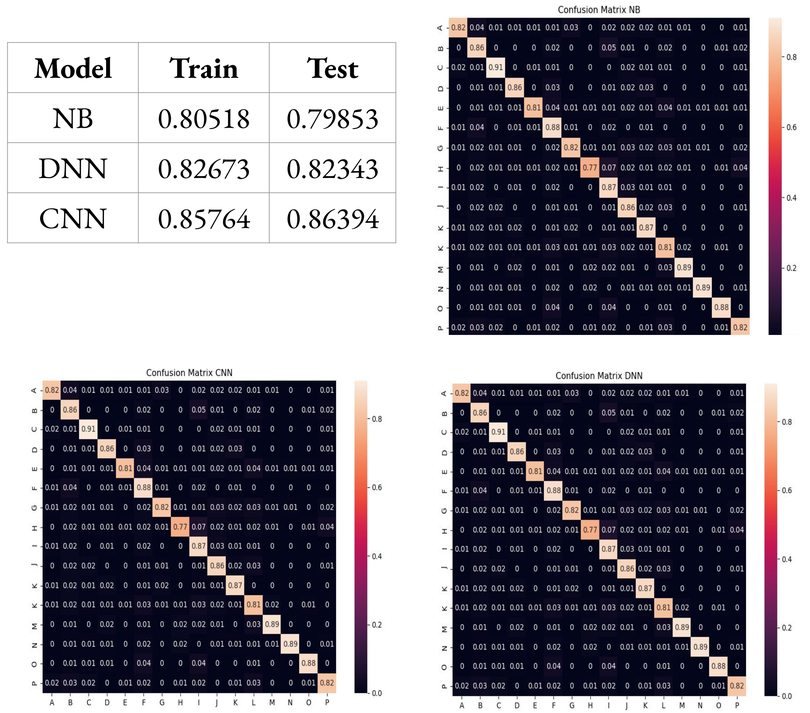
The dense network consisted of 3 hidden layers with 6, 4, and 2 times the number of classes neurons in them respectively. The hidden layers have the ReLU function as activation and the output layer has softmax as the activation function.

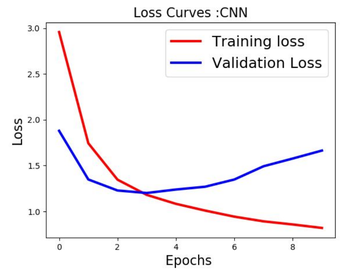
For the CNN there were 3 convolutional layers with padding before the final fully connected one. The architecture was inspired by the one described in [3]. For both the DNN and CNN batch normalization was done after each activation.

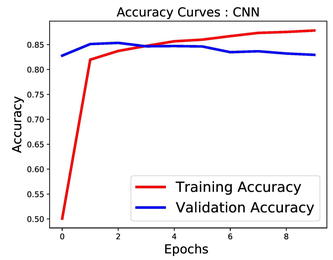
$$a^{[j]} = f\left(\sum_i w_i a_i^{[j-1]} + b\right) \qquad \hat{\theta}_{yi} = \frac{\sum_{x \in T} x_i + \alpha}{\sum_{i=1}^{n} \sum_{x \in T} x_i + \alpha n}$$

Results

Model	Train	Test
NB	0.80518	0.79853
DNN	0.82673	0.82343
CNN	0.85764	0.86394







Discussion

The conclusions that can be made from this project is that text theme classification can be done. A significant increase in performance can be achieved if the classes are compounded into more general ones. However, the results for more specific classes proved itself to be more complicated.
Interesting to note is that the more complicated models did not perform severely much better than the more crude naive bayes model. We were expecting the more complicated models to perform better as the CNN has been made to work on text processing. This could be due to dataset as the quality of it might not be the best for the purpose of this project. An indication of this is that we achieved much better result by changing just some of the labels.

Future steps

Future work on this project consists of further tuning of the CNN, testing a RNN, review and find another dataset and also train a model for sentence selection.

References

[1] Kasra Madadipouya. Csv dataset of 76,000 quotes, suitable for quotes recommender systems or other analysis., 07 2016.
[2] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. 2014.
[3] Akshat Maheshwari. Report on text classification using cnn, rnn han. https://medium.com/jatana, 2018.