CSC411: Assignment 3 Bonus

Due on Saturday, March 24, 2018

Yu-Chien Chen, Hunter Richards

March 25, 2018

Problem 1

a) The purpose of the bonus project is to build a better fake news detector. The real news headlines were collected from The New York Times, The Guardian and the ones provided for project 3. The fake news headlines were from the same website used for project 3. Instead of only collecting trump related headlines, we decided to broaden the scope to include world news. Both The New York Times and The Guardian data were harvested using a script found on github[1]. The real and fake headlines were then cleaned using the modified version of clean_script.py (script provided for project 3). There are now 4736 headlines for both real news and fake news.

An attempt was made to use a convolutional neural net to classify news headlines. It uses an embedding layer, two convolution layers, ReLU activation, dropout, and a final fully connected layer. The embedding layer is purposed with reducing the very high dimensional input to a smaller, fixed vector size. The vocabulary is much larger this time so one hot vectors are much larger and slow computation significantly. The convolution layers are meant to look at adjacent words. More meaning can be derived from a sequence of words rather than looking at words individually. Finally, no max pool layers were used, which are typical of a convolution net, because of how they add invariance to the output. We want the model to detect the ordering of words because this may be important in a news headline. Adding max pool layers will work against this.

The CNN model did not perform as well as the Logistic Regression model from project 3. It consistently performed $\sim 10\%$ worse on the validation set. Therefore, the Logistic Regression model was chosen because of its superior performance. After re-tuning the hyperparameters with a grid search, the model produced the following learning curve on the new dataset:

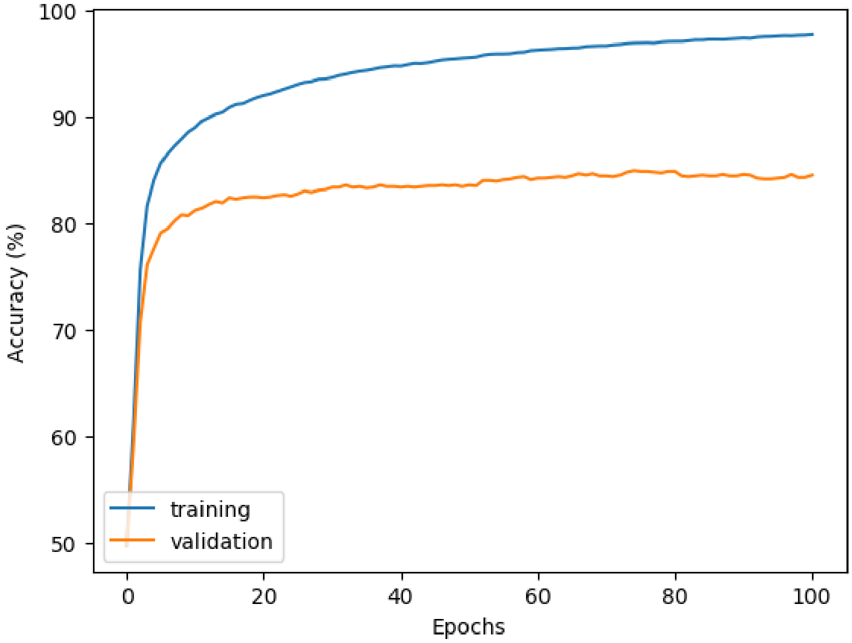

Figure 1

[1] https://github.com/genyunus/Detecting_Fake_News/tree/master/Scraping_Data

The model scored an accuracy of 97.71%, 84.51%, and 84.44% on the new training, validation, and test set, respectively. The performance on the test set from project 3 was used to evaluate the improvement of the model. It scored 87.53%, an increase of 3% over the previous Logistic Regression implementation, 5% over Naive Bayes, and 10% over the Decision Tree. This is not the substantial increase we hoped for, unfortunately.

b) Because we are still using the Logistic Regression model from project 3, we can look at the top 10 θ's and their corresponding words again.

The top ten positive θ's and their corresponding words were:

```
[(1.3919568, ’comment’), (1.232754, ’soros’), (1.2294954, ’breaking’),
(1.2014278, ’dakota’), (1.2011237, ’aleppo’), (1.1951044, ’hillary’),
(1.1802721, ’podesta’), (1.1422879, ’wikileaks’), (1.1061746, ’clintons’),
(1.0937072, ’doj’)]
```

The top ten negative θ's and their corresponding words were:

```
[(-1.7718631, ’USA’), (-1.4721, ’pictures’), (-1.3230065, ’tariffs’),
(-1.2287176, ’turnbull’), (-1.149365, ’role’), (-1.1277608, ’mueller’),
(-1.1214317, ’tillerson’), (-1.0836383, ’cup’),
(-1.0504378, ’charlottesville’), (-1.0279093, ’bid’)]
```

Note there are no stop-words in either list. We can see that the most decisive words have changed drastically from project 3, where the word ”trump” was commonly seen. It does not appear at all here. This is likely because the previous dataset was vetted such that it only contained headlines with ”trump”. This time that bias was removed from the dataset.