

CMPS 242 - Fall 17

Homework 5 - Report

Nursultan Kabylkas
Ramesh Jayaraman

11/17/2017

1. Introduction

The objective of this project is to classify tweets, tweeted by Hillary Clinton and Donald Trump. This was achieved by implementing a LSTM Recurrent Neural Net, using tensor flow, training it with the data provided, and tested using the test data set provided. The results from the test set were submitted to Kaggle, to rank different implementations by their accuracy.

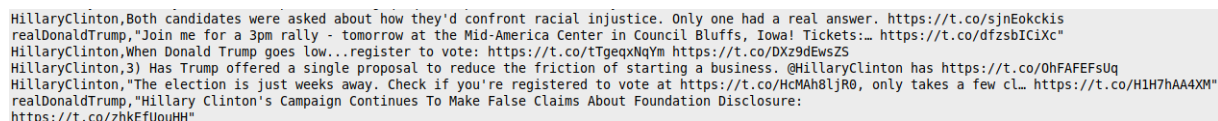
This report is structured to provide an introduction to the work done, followed by the details of the implementation, and then the results that were obtained.

2. Implementation

In this section, we explain all aspects of our implementation, starting with the characteristics of the input data set, how we perform pre-processing, then tokenize the data. We then give details about the implementation of our Recurrent Neural Net, and how we train it.

2.1 Data set characteristics

The input data set has the format as shown in Figure 1. The input file was formatted as a comma-separated values (".csv") file, where the different fields are separated by commas.



```
HillaryClinton,Both candidates were asked about how they'd confront racial injustice. Only one had a real answer. https://t.co/sjnEokckis
realDonaldTrump,"Join me for a 3pm rally - tomorrow at the Mid-America Center in Council Bluffs, Iowa! Tickets: https://t.co/dfzsbICiXc"
HillaryClinton,When Donald Trump goes low...register to vote: https://t.co/tTgeqNqYm https://t.co/DXz9dEwsZS
HillaryClinton,(3) Has Trump offered a single proposal to reduce the friction of starting a business. @HillaryClinton has https://t.co/0hFAFEFSUq
HillaryClinton,"The election is just weeks away. Check if you're registered to vote at https://t.co/HcMAh8ljR0, only takes a few cl... https://t.co/H1H7hAA4XM"
realDonaldTrump,"Hillary Clinton's Campaign Continues To Make False Claims About Foundation Disclosure: https://t.co/zhkEfUouHH"
```

Figure 1: This image shows the format of the input file.

The first field is the Twitter handle of the candidate, which serves as the label for the tweet, HillaryClinton (Hillary Clinton),realDonaldTrump (Donald Trump). The second field contains the tweet. There were a total of 4743 tweets in the input data set.

The test data set was given in the same format, ie. a .csv file, containing a total of 1701 tweets, but had the handle replaced with "none".

As we could not validate our predictions due to the absence of labels in the test data set, the results needed to be submitted to kaggle, as a .csv file, formatted as shown in Figure 2, to be compared against other groups.

[tweet id],[probability of Donald Trump],[probability of Hillary Clinton]

Figure 2: Submission Format

The next section provides more information about the preprocessing that was performed on the dataset.

2.2 Preprocessing

2.3 Tokenizing

Talk about how we tokenized. lead into rnn section

2.4 Validation

For this experiment we used Single Split method for cross validation. 20% of the data points were randomly selected, extracted from the data set and kept for further validation. Remaining 80% of the data points were actually used for training.

2.5 Hyperparameter selection

The hyperparameters that we played with are batch size, state vector size, and the number of epochs spent to train the model. It was noticed that the batch size of 1 gives the best results regardless of the chosen model. The state vector size was x^2 where $x \in [1, 2, \dots, 9]$. We noticed that the model stops improving $x > 70$, and it becomes computationally infeasible given the time frame we were given. It was mentioned in the class that one of the methods of regularization is "early training stopping". For this reason, we ran the program for 20 epochs for all values of x . It was also interesting to notice that batch size of 1 made the best results to occur in early epochs. We claim that this is due to the faster convergence. Finally, we picked the best accuracy on validation set, and re-run the training with the best x and upto the chosen epoch.

2.5 RNN implementations

Different versions of RNN models were implemented, trained and tested. Namely, the following three RNN methods were tried out:

- Vanilla RNN
- LSTM
- GRU

2.5.1 Vanilla RNN

talk about training, and then lead into testing and results.

3.Results

Give intro to what we ran, config etc.

3.1 rnn results

give results from rnn

3.2 Compare with logistic regression

Compare it with logistic.

4. Conclusion

Conclude work.