Homework 5 Report

Jae Young Kim

Applied Data Science, University of Southern California,

Los Angeles, California 90089, USA

(Dated: April 8, 2021)

Abstract

Through this report, ps5_tweets_labels.csv data set was used to classify the tweets into 5 classes based on their sentiment. Naive Bayes Classifier and LSTM model were used to train the data. To put adequate input to the models, number of words and labels were counted and pretrained word embedding model, "nnlm-en-dim128" was loaded. Finally, 0.474 and 0.591 accuracy were achieved by each model.

I. INTRODUCTION

To make a classifier which classifies tweets in terms of their sentiment, a Naive Bayes classifier was trained. 37042 tweets with numerical label from 1 to 5 was given. To calculate the posterior probability of each category, preprocessing was done for the given data set. The given tweets were tokenized and normalized with nltk library. With the preprocessed data, train and test set were split. By creating dictionary of the normalized words and sentiment, it was possible to calculate the likelihood of the words.

Fianlly, Naive Bayes Classifier was trained with prior and likelihood and LSTM model was trained with pretrained word2vec model. The accuracy of the trained model on test set was 0.474 and 0.591 for each. However, the confusion matrix showed that most data were concentrated near diagonal line, which means the trained Naive Bayes classifier and LSTM model both worked well.

II. DATA PREPROCESSING

To preprocess data, nltk library was used. Special characters and web addresses were deleted, stop words were removed and stemming was implemented.

```
Original Text :  Alert: Largest wholesale and retail Ongata Rongai Kware
Open Air Market faces imminent closure this day in a mid to combat covid
19 spread. This move will leave consumers at the mercy of local retailer
s who will enjoy freedom of setting prices.
Be on the lookout!
#GikombaCorona
###################################
Normalized Text :  ['alert', 'largest', 'wholesal', 'retail', 'ongata',
'rongai', 'kware', 'open', 'air', 'market', 'face', 'immin', 'closur',
'day', 'mid', 'combat', 'covid', '19', 'spread', 'move', 'leav', 'consu
m', 'merci', 'local', 'retail', 'enjoy', 'freedom', 'set', 'price', 'loo
kout', 'gikombacorona']
```

FIG. 1. Example of input and output of text normalizer

Figure 1 shows an example of the input and output of text normalizer.

III. MODEL SELECTION AND EVALUATION

To train a classifier, Naive Bayes Classifier and LSTM model was trained. 10-fold cross validation was used to train and test the performance of Naive Bayes Classifier but it was not used to LSTM model since it takes too much time to train LSTM model. Alternatively, 15% of validation set and 15% of test set were selected from the given data set. "nnlm-en-dim128" model was used as a word embedding model. The model is a token based text embedding trained on English Google News 200B corpus. Figure 2 shows the architecture of trained LSTM model nad the following table shows the performance of the model.

```
Model: "sequential_10"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_10 (LSTM)               (None, 32)                20608
_________________________________________________________________
dense_11 (Dense)             (None, 5)                 165
=================================================================
Total params: 20,773
Trainable params: 20,773
Non-trainable params: 0
_________________________________________________________________
```

FIG. 2. Architecture of LSTM model

Model	Naive Bayes Classifier	LSTM model
Accuracy	0.474	0.591
Precision	0.525	0.617
Recall	0.462	0.594
F1 Score	0.478	0.603

Niave Bayes Classifier showed 0.474 accuracy and LSTM model showed 0.591 accuracy. Figure 3 and Figure 4 shows the heat map of confusion matrix of the models. Figure 4 shows darker color in diagonal line and it means LSTM model showed better result.

IV. CONCLUSIONS

To sum up, to classify given tweet data into 5 groups, Extremely, Negative, Negative, Neutral, Positive and Extremely Positive, Naive Bayes Classifier and LSTM model were trained. To put proper input to the models, number of words and labels were counted and pretrained word embedding model was loaded. Finally, the two models showed 0.474 and 0.591 accuracy for each.

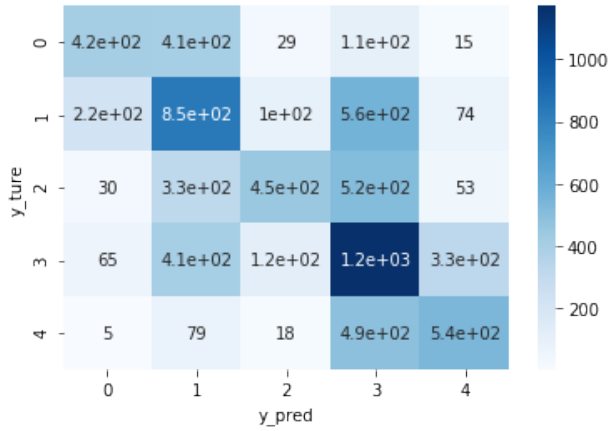

FIG. 3. Confusion matrix of Naive Bayes classifier

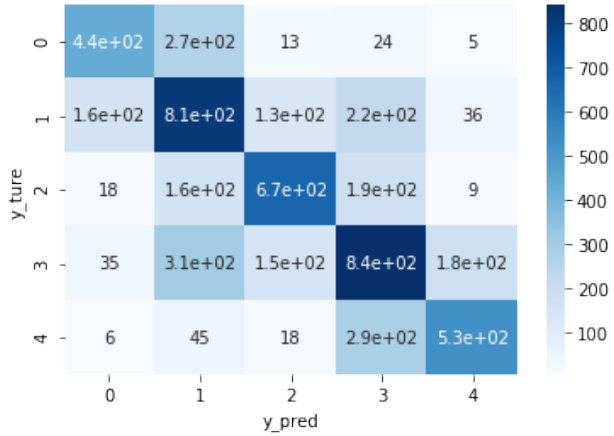

FIG. 4. Confusion matrix of LSTM model

DATA AVAILABILITY

Data is available at
https://github.com/USC-DSCI-552-Spring2021/dsci552-spring2021-32416d-ps5-jeayoung114/data

CODE AVAILABILITY

Code is available at
https://github.com/USC-DSCI-552-Spring2021/dsci552-spring2021-32416d-ps5-jeayoung114/code/hw5