DEEP LEARNING FOR NATURAL LANGUAGE PROCESSING

CODING ASSIGNMENT REPORT

Suba Shree V S
18 5001 171

PROBLEM STATEMENT:

Process the training data consisting of various recent themes centered around current events. There are a series of sentiments given in the training data. Predict the sentiment of the test data.

OVERVIEW OF CODE:

LIBRARIES USED:

- Pandas
- Numpy
- Matplotlib
- Scikitlearn
- preprocessor, tweet-preprocessor, re (regular expression library)

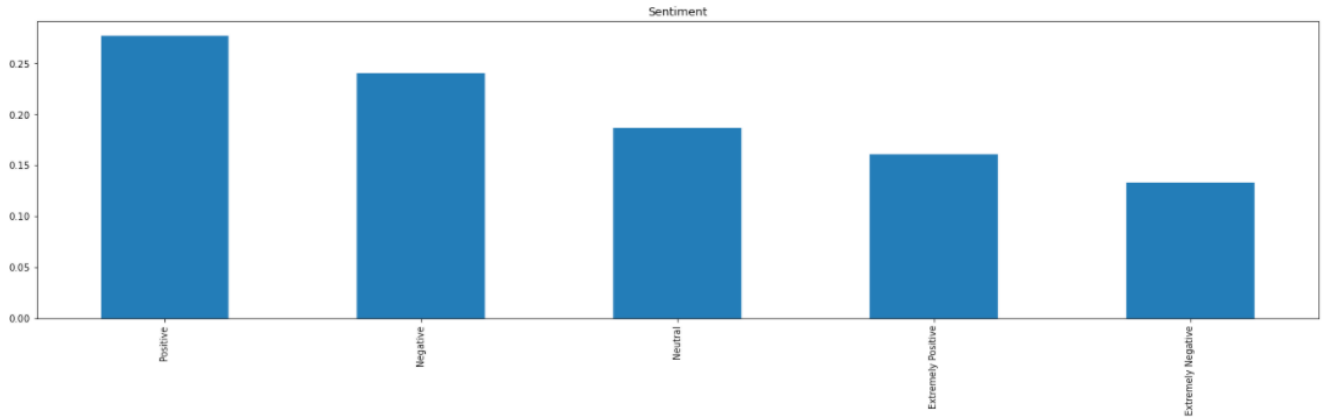
DATASETS:

1. **train.csv**
   Features: UserName, ScreenName, Location, TweetAt, OriginalTweet, Sentiment
   Number of rows: 41157

2. **test.csv**
   Features: UserName, ScreenName, Location, TweetAt, OriginalTweet
   Number of rows: 3798

DATA ANALYSIS:

Distribution of Tweet Sentiments:



Distribution of various locations:

```
train['Location'].value_counts()
```

```
London                          540
United States                   528
London, England                 520
New York, NY                    395
Washington, DC                  373
                               ...
Jackson Hole, WY                  1
The City of London                1
Milton keynes , England           1
Saratoga Springs, NY              1
Anywhere There's Internet         1
Name: Location, Length: 12220, dtype: int64
```

DATA CLEANING & PRE-PROCESSING:

Removed special characters such as hashtags, mentions, numbers and punctuations from the OriginalTweet data using the regular expression library.

Function to clean the dataset (combining tweet_preprocessor and regular expression):

```
import preprocessor as p

# custum function to clean the dataset (combining tweet_preprocessor and reguar expression)
def clean_tweets(df):
  tempArr = []
  for line in df:
    # send to tweet_processor
    tmpL = p.clean(line)
    # remove puctuation
    tmpL = REPLACE_NO_SPACE.sub("", tmpL.lower()) # convert all tweets to lower cases
    tmpL = REPLACE_WITH_SPACE.sub(" ", tmpL)
    tempArr.append(tmpL)
  return tempArr
```

Cleaned vs Uncleaned Tweets:

1. Train Dataset

OriginalTweet	Sentiment	clean_tweet
@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Neutral	and and
advice Talk to your neighbours family to excha...	Positive	advice talk to your neighbours family to excha...
Coronavirus Australia: Woolworths to give elde...	Positive	coronavirus australia woolworths to give elder...
My food stock is not the only one which is emp...	Positive	my food stock is not the only one which is emp...
Me, ready to go at supermarket during the #COV...	Extremely Negative	me ready to go at supermarket during the outbr...
As news of the regionÂ□s first confirmed COVID...	Positive	as news of the regions first confirmed covid 1...
Cashier at grocery store was sharing his insig...	Positive	cashier at grocery store was sharing his insig...
Was at the supermarket today. Didn't buy toile...	Neutral	was at the supermarket today didnt buy toilet ...
Due to COVID-19 our retail store and classroom...	Positive	due to covid 19 our retail store and classroom...
For corona prevention,we should stop to buy th...	Negative	for corona preventionwe should stop to buy thi...

2. Test Dataset

OriginalTweet	clean_tweet
You never eaten the pigs cat dog or food from ...	you never eaten the pigs cat dog or food from ...
@calebmealer @thebradfordfile @realDonaldTrump...	very true china has done a great job of more t...
Even though the Law Library is closed, ALL sub...	even though the law library is closed all subs...
With Gov Hogan's announcement that all bars, r...	with gov hogans announcement that all bars res...
@RicePolitics @MDCounties Craig, will you call...	craig will you call on the general assembly to...
Meanwhile In A Supermarket in Israel -- People...	meanwhile in a supermarket in israel people...
Did you panic buy a lot of non-perishable item...	did you panic buy a lot of non perishable item...
Asst Prof of Economics @cconces was on @NBCPhi...	asst prof of economics was on talking about he...
Gov need to do somethings instead of biar je r...	gov need to do somethings instead of biar je r...
I and @ForestandPaper members are committed to...	i and members are committed to the safety of o...

METHODOLOGY (Support Vector Classification)

VECTORIZATION:

```
from sklearn.feature_extraction.text import CountVectorizer

# vectorize tweets for model building
vectorizer = CountVectorizer(binary=True, stop_words='english')

# learn a vocabulary dictionary of all tokens in the raw documents
vectorizer.fit(list(x_train) + list(x_test))

# transform documents to document-term matrix
x_train_vec = vectorizer.transform(x_train)
x_test_vec = vectorizer.transform(x_test)
```

```
<28809x40739 sparse matrix of type '<class 'numpy.int64'>'
        with 402974 stored elements in Compressed Sparse Row format>
```

MODEL BUILDING USING SVC:

```
# classify using support vector classifier
svm = svm.SVC(kernel = 'linear', probability=True)
```

```
# fit the SVC model based on the given training data
prob = svm.fit(x_train_vec, y_train).predict_proba(x_test_vec)
```

```
# perform classification and prediction on samples in x_test
y_pred_svm = svm.predict(x_test_vec)
```

```
from sklearn.metrics import accuracy_score
print("Accuracy score for SVC is: ", accuracy_score(y_test, y_pred_svm) * 100, '%')
```

```
Accuracy score for SVC is:  63.64593456430191 %
```

```
y_pred_svm
```

```
array([4, 1, 3, ..., 4, 3, 3])
```

Accuracy Obtained: 63.65
