

```
[2]: #Basic Imports
import pandas as pd
import numpy as np

In [3]: #sklearn imports
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn import feature_extraction, model_selection, naive_bayes, pipeline, manifold, preprocessing
from sklearn.dummy import DummyClassifier
from sklearn.feature_extraction.text import TfidfVectorizer, CountVecorizer

In [4]: from keras.layers import LSTM,Dense,Dropout,Embedding,CuDNNLSTM,Bidirectional

In [5]: from sklearn import BayesSearchCV
from sklearn.model_selection import GridSearchCV
from sklearn.space import Real, Categorical, Integer
from sklearn.metrics import beta_score, make_scorer
from keras.preprocessing.sequence import pad_sequences
from keras.callbacks import EarlyStopping
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
from matplotlib import pyplot
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score

In [6]: from tensorflow import keras

In [7]: import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud
from tqdm import tqdm

In [10]: import sys
sys.path.insert(0, '/Users/seanyboy/Documents/Flatiron/Phase_4/Twitter_Sentiment_Project/Helper_Functions')
import functions as fn

[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data] /Users/seanyboy/nltk_data...
[nltk_data] Package averaged_perceptron_tagger is already up-to-
[nltk_data] date!

In [11]: from tensorflow.keras import Sequential
from tensorflow.keras import layers

In [12]: import nltk
from nltk.corpus import stopwords, wordnet
from nltk.tokenize import regep_tokenize, word_tokenize, RegexpTokenizer
from nltk.stem import WordNetLemmatizer
from nltk import word_tokenize
from gensim.models import word2vec
from nltk import pos_tag
from keras.preprocessing.text import Tokenizer
```

Import Dataset with Additional Negative Examples

```
In [64]: new_df = pd.read_csv('additional_negs')

In [65]: new_df['processed_text'] = fn.ProcessTweet(new_df['tweet_text'],remove HTML=True,replace_moji_bake=True,strip_punct=True)

In [66]: #Combining Neutral and Positive Values
new_df['Sentiment']=new_df['Sentiment'].replace({"Positive": 0, "Negative": 1, "Neutral": 3}).astype(np.uint8)
#Setting the Minority Class Negative as the Value of "1"
new_df['Sentiment']=new_df['Sentiment'].replace(3,0).copy()

In [67]: # new_df = new_df[new_df['Sentiment']!='Neutral']

In [68]: new_df['Sentiment'].value_counts(normalize=True)

Out[68]:
0    0.823831
1    0.176169
Name: Sentiment, dtype: float64
```

Tokenizing and Lemmatizing

At this step I will again perform the same lemmatization performed on the traditional models. I will not use the same pipeline as it creates TFIDF vectors. I will be using the pretrained GLOVE vectors.

```
In [69]: #Import Stopwords
sw = stopwords.words('english')
len(sw)

Out[69]: 179

There are a 179 stopwords in this list, I may experiment with adding stopwords at some point.
```

```
In [70]: len(sw)

Out[70]: 179

In [71]: from keras.preprocessing.text import Tokenizer
t = Tokenizer()

In [72]: from sklearn.feature_extraction.text import TfidfVectorizer

In [73]: tokenizer = TfidfVectorizer.build_tokenizer

In [74]: new_df['processed_text']

Out[74]:
0      I have a three G iPhone. After three hrs tw...
1      Know about ? Awesome iPad/iPhone app that yo...
2      Can not wait for #iPad two also. They shoul...
3      I hope this year's festival is not as crashy ...
4      great stuff on Fri #SXSW: Marissa Mayer (Goog...
...
10150  RT : Thinking of upgrading to #osennite? Think...
10151      why is not group facetime a thing wtf
10152  Being held hostage at - They are replacing th...
10153  hey is it normal for my laptop charger to be ...
10154  My iPhone five 's photos are no longer downlo...
Name: processed_text, Length: 10155, dtype: object

In [75]: new_df

Out[75]:
```

I will take advantage of the doc preparer function that takes in a the words from the dataset and replaces any remaining punctuation, removes stopwords, and lemmatizes the words in the dataset.

```
In [77]: def doc_preparer(doc, stop_words=sw):
'''
:param doc: a document from a corpus
:return: a document string with words which have been
        lemmatized,
        parsed for stopwords,
        made lowercase,
        and stripped of punctuation and numbers.
'''
    regex_token = RegexpTokenizer(r'[a-zA-Z]+(?:' + '[a-z]+)?')
    doc = regex_token.tokenize(doc)
    doc = [word.lower() for word in doc]
    doc = [word for word in doc if word not in sw]
    # doc = pos_tag(doc)
    doc = [(word[0], get_wordnet_pos(word[1])) for word in doc]
    # lemmatizer = nltk.stem.wordnet.WordNetLemmatizer()
    # doc = lemmatizer.lemmatize(word[0], word[1]) for word in doc
    return doc

In [78]: # new_df['processed_text']=doc_preparer(doc, sw) for doc in new_df['processed_text']

In [79]: new_df.head(5)
```

```
Out[79]:
```

	tweet_text	Product	Sentiment	processed_text
0	.@wesley83 I have a 3G iPhone. After 3 hrs tw...	iPhone	1	I have a three G iPhone. After three hrs t...
1	@jessedee Know about @fudapp ? Awesome iPad/...	iPad or iPhone App	0	Know about ? Awesome iPad/iPhone app that yo...
2	@swonderlin Can not wait for #iPad 2 also. The...	iPad	0	Can not wait for #Pod two also. They shoul...
3	@sxswv I hope this year's festival isn't as cra...	iPad or iPhone App	1	I hope this year's festival is not as crashy ...
4	@sxstatoe great stuff on Fri #SXSW: Marissa M...	Google	0	great stuff on Fri #SXSW: Marissa Mayer (Goog...

```
In [80]: new_df.info()

Out[80]:
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10155 entries, 0 to 10154
Data columns (total 4 columns):
#   Column      Non-Null Count  Dtype
---  --
0   tweet_text  10155 non-null  object
1   Product     4501 non-null  object
2   Sentiment   10155 non-null  uint8
3   processed_text  10155 non-null  object
dtypes: object(3), uint8(1)
memory usage: 248.0+ KB
```

Adding Glove Word Embeddings

Train/Test Split

```
In [81]: new_df['processed_text']

Out[81]:
0      I have a three G iPhone. After three hrs t...
1      Know about ? Awesome iPad/iPhone app that yo...
2      Can not wait for #iPad two also. They shoul...
3      I hope this year's festival is not as crashy ...
4      great stuff on Fri #SXSW: Marissa Mayer (Goog...
...
10150  RT : Thinking of upgrading to #osennite? Think...
10151      why is not group facetime a thing wtf
10152  Being held hostage at - They are replacing th...
10153  hey is it normal for my laptop charger to be ...
10154  My iPhone five 's photos are no longer downlo...
Name: processed_text, Length: 10155, dtype: object

In [82]: X = new_df['processed_text']
y = new_df['Sentiment']

In [83]: len(X)

Out[83]: 10155

In [84]: #Splitting First into a Train Set and Remaining Data Set
X_train, X_rem, y_train, y_rem = train_test_split(X,y, train_size=0.8,random_state=42,stratify=y)

In [85]: #Splitting the Remaining Data into Validation and Test Data
X_val, X_test, y_val, y_test = train_test_split(X_rem,y_rem, test_size=0.5,random_state=42,atratify=y_rem)

In [86]: len(X_val)

Out[86]: 1015
```

I have split the data twice, the holdout data will be the final data to test on for all the models that I will attempt.

Tokenizing for Glove Model

```
In [87]: X_train_tokens=[doc_preparer(doc, sw) for doc in X_train]
X_val_tokens = [doc_preparer(doc, sw) for doc in X_val]
X_test_tokens = [doc_preparer(doc,sw) for doc in X_test]

In [88]: X_train_tokens
```



```
'conference',
'launch',
'apple',
'make',
'stop',
['rt', 'google', 'asks', 'want', 'know'],
'leaving',
'excellent',
'cancel',
'writing',
'compelling',
'conchords',
'android',
'dev',
'newcup',
'sxsw1',
['cnet',
'apple',
'schooling',
'schooling',
'marketing',
'execs',
'pop',
'ipad',
'cnet'],
['launch'],
['rt', 'one', 'fav', 'photos', 'sxsw', 'far', 'google', 'sxsw'],
'cancel',
'turn',
'push',
'apple',
'fetch',
'manual',
'nodes',
'email',
'mail',
'conchords',
'cal',
'fetch',
'new',
'data',
'saves',
'updates',
'iphone',
'sxsw1',
['sponsored', 'apple', 'plaid', 'boys', 'attracted'],
'smartphone',
'breakdown',
'official',
'unofficial',
'ninety',
'apple',
'five',
'iphones',
'rt',
'android',
'three',
'blackberry',
'two',
'symbian',
'windows',
['boot',
'rt',
'blog',
'posts',
'culture',
'mobile',
'sxsw',
'updates',
'iphone',
'ip'],
['waiting', 'google', 'keynote', 'start', 'sxsw'],
'sxsw',
'already',
'apple',
'happening',
'google',
'rt',
'refine',
'searches',
'protect',
'smst',
['rt',
'new',
'ubersocial',
'iphone',
'apple',
'store',
'includes',
'uberguide',
'sxsw',
'sponsored',
'scheduling',
'informal',
'unscientific',
'competition',
'apple',
'far',
'computer',
'choice',
'sxsw',
'iphone',
'contest'],
['rt',
'new',
'came',
'travolta',
'rt',
'broken',
'arrow',
'sxsw',
'thought',
'two',
'first',
'apple',
'rt',
'rt',
'rt',
'google',
'launch',
'major',
'new',
'social',
'network',
'called',
'circles',
'possibly',
'code',
'sxsw1'],
['letting',
'interactive',
'file',
'closing',
'mail',
'opening',
'party',
'badcase',
'yousareinteractive',
'peopleilovingout',
'sxsw',
['today',
'last',
'apple',
'pop',
'sxsw',
'vacation',
'starts',
'computer',
'great',
'experience'],
['conting',
'new',
'iphone',
'apple',
'app',
'sxsw',
'calendar',
'year',
'makes',
'phone',
'waterproof',
'light',
'codebases'],
['rt',
'rt',
'google',
'launch',
'major',
'new',
'social',
'network',
'called',
'circles',
'possibly',
'code',
'sxsw1'],
['ick', 'yo', 'autocorrect', 'ruse'],
'one',
'iphone',
'five',
'conferences',
'stopped',
'working',
'code',
'send',
'apple',
'forgive',
'hope',
'apple',
'approves',
'relay',
'apple',
'network',
'mobile',
'phone',
'sharing',
'app',
'quizzes',
'easier',
'share',
'google',
'download'],
['still', 'cannot', 'merge', 'multiple', 'apple', 'id', 'fail'],
'apple',
'open',
'popup',
'apple',
'friday',
'six',
'code',
'congress',
'sxsw',
'rt',
'satesman',
'com'],
['rt',
'attending',
'sxsw',
'apple',
'guide',
'free',
'apple',
'download',
'itunes',
'ip',
'new',
'ubersocial',
'iphone',
'store',
'includes',
'uberguide',
'sxsw',
'sponsored',
'calendar',
['spilled',
'beans',
'beans',
'platform',
'flipboard',
'iphone',
'team',
'stated',
'calendar',
'sxflip',
'sxsw',
'gr',
'right'],
'apple',
'post',
'twitter',
'apple',
'instagram',
'fb',
'shutterfly',
'apple',
'sfly',
'shutdown'],
'apple',
'best',
'conference',
'apple',
'ever',
'google',
'google',
'four',
'sqchat'],
'new',
'inspiration',
'sabi',
'looking',
'spanish',
'speaking',
'trend',
'acout',
'apple',
'austin',
'texas',
'code',
'back',
'new',
'problems',
'something',
'right'],
'apple',
'later',
'reilly',
'line',
'door',
'sxsw',
'apple',
'store',
'ipad',
'two',
'new',
'posts',
'apple',
'iphone',
'apps',
'scheduling',
'south',
'southwest',
'apple',
'parts = line.split()
word = parts[0].decode('utf-8')
if word in total_vocabulary:
    vector = np.array(parts[1:], dtype=np.float32)
    embedding_vector[word] = vector

In [96]: #Creating Sequence Tokens
token = Tokenizer()
token.fit_on_texts(X_train)
seq = token.texts_to_sequences(X_train)

In [97]: X_train_tokens[0:4]

Out[97]:
['sxsw',
'mistakes',
'made',
'building',
'netflix',
'iphone',
'glad',
'see',
'source',
'code'],
['great',
'visualisation',
'ghost',
'movement',
'logic',
'apple',
'doodles',
'apple',
'sxsw',
'details'],
['iphone',
'worries',
'drop',
'new',
'venues',
'talk',
'convergence',
'take',
'case',
'sxsw',
'sxsw1'],
['heard',
'downtown',
'location',
'sold',
'one',
'merchandise',
'last',
'apple',
'one',
'two']],
In [98]: seq[0:4]
```


