Baseline Models for Twitter Sentiment Classification

In this notebook, we will test the effectiveness of non-deep learning models. We will work with Naive Bayes.

In [1]:
```
import pandas as pd
import re
import nltk
import string
import os
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from nltk.tokenize import word_tokenize, sent_tokenize
from nltk.stem.wordnet import WordNetLemmatizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from keras.preprocessing.text import Tokenizer
from sklearn.model_selection import train_test_split
import glob, os
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Embedding, Dense
import numpy as np
from sklearn.feature_extraction.text import CountVectorizer, TfidfTran
sformer
from sklearn.metrics import confusion_matrix, classification_report, a
ccuracy_score, f1_score
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras.preprocessing.text import Tokenizer
```

```
/Users/shivaomrani/opt/anaconda3/envs/neural_networks/lib/python3.7/
site-packages/tensorflow/python/framework/dtypes.py:516: FutureWarni
ng: Passing (type, 1) or '1type' as a synonym of type is deprecated;
in a future version of numpy, it will be understood as (type, (1,))
/ '(1,)type'.
  _np_qint8 = np.dtype([("qint8", np.int8, 1)])
/Users/shivaomrani/opt/anaconda3/envs/neural_networks/lib/python3.7/
site-packages/tensorflow/python/framework/dtypes.py:517: FutureWarni
ng: Passing (type, 1) or '1type' as a synonym of type is deprecated;
in a future version of numpy, it will be understood as (type, (1,))
/ '(1,)type'.
  _np_quint8 = np.dtype([("quint8", np.uint8, 1)])
/Users/shivaomrani/opt/anaconda3/envs/neural_networks/lib/python3.7/
site-packages/tensorflow/python/framework/dtypes.py:518: FutureWarni
```

```
ng: Passing (type, 1) or '1type' as a synonym of type is deprecated;
in a future version of numpy, it will be understood as (type, (1,))
/ '(1,)type'.
  _np_qint16 = np.dtype([("qint16", np.int16, 1)])
/Users/shivaomrani/opt/anaconda3/envs/neural_networks/lib/python3.7/
site-packages/tensorflow/python/framework/dtypes.py:519: FutureWarni
ng: Passing (type, 1) or '1type' as a synonym of type is deprecated;
in a future version of numpy, it will be understood as (type, (1,))
/ '(1,)type'.
  _np_quint16 = np.dtype([("quint16", np.uint16, 1)])
/Users/shivaomrani/opt/anaconda3/envs/neural_networks/lib/python3.7/
site-packages/tensorflow/python/framework/dtypes.py:520: FutureWarni
ng: Passing (type, 1) or '1type' as a synonym of type is deprecated;
in a future version of numpy, it will be understood as (type, (1,))
/ '(1,)type'.
  _np_qint32 = np.dtype([("qint32", np.int32, 1)])
/Users/shivaomrani/opt/anaconda3/envs/neural_networks/lib/python3.7/
site-packages/tensorflow/python/framework/dtypes.py:525: FutureWarni
ng: Passing (type, 1) or '1type' as a synonym of type is deprecated;
in a future version of numpy, it will be understood as (type, (1,))
/ '(1,)type'.
  np_resource = np.dtype([("resource", np.ubyte, 1)])
/Users/shivaomrani/opt/anaconda3/envs/neural_networks/lib/python3.7/
site-packages/tensorboard/compat/tensorflow_stub/dtypes.py:541: Futu
reWarning: Passing (type, 1) or '1type' as a synonym of type is depr
ecated; in a future version of numpy, it will be understood as (type
, (1,)) / '(1,)type'.
  _np_qint8 = np.dtype([("qint8", np.int8, 1)])
/Users/shivaomrani/opt/anaconda3/envs/neural_networks/lib/python3.7/
site-packages/tensorboard/compat/tensorflow_stub/dtypes.py:542: Futu
reWarning: Passing (type, 1) or '1type' as a synonym of type is depr
ecated; in a future version of numpy, it will be understood as (type
, (1,)) / '(1,)type'.
  _np_quint8 = np.dtype([("quint8", np.uint8, 1)])
/Users/shivaomrani/opt/anaconda3/envs/neural_networks/lib/python3.7/
site-packages/tensorboard/compat/tensorflow_stub/dtypes.py:543: Futu
reWarning: Passing (type, 1) or '1type' as a synonym of type is depr
ecated; in a future version of numpy, it will be understood as (type
, (1,)) / '(1,)type'.
  _np_qint16 = np.dtype([("qint16", np.int16, 1)])
/Users/shivaomrani/opt/anaconda3/envs/neural_networks/lib/python3.7/
site-packages/tensorboard/compat/tensorflow_stub/dtypes.py:544: Futu
reWarning: Passing (type, 1) or '1type' as a synonym of type is depr
ecated; in a future version of numpy, it will be understood as (type
, (1,)) / '(1,)type'.
  _np_quint16 = np.dtype([("quint16", np.uint16, 1)])
/Users/shivaomrani/opt/anaconda3/envs/neural_networks/lib/python3.7/
site-packages/tensorboard/compat/tensorflow_stub/dtypes.py:545: Futu
reWarning: Passing (type, 1) or '1type' as a synonym of type is depr
ecated; in a future version of numpy, it will be understood as (type
```

```
, (1,)) / '(1,)type'.
  _np_qint32 = np.dtype([("qint32", np.int32, 1)])
/Users/shivaomrani/opt/anaconda3/envs/neural_networks/lib/python3.7/
site-packages/tensorboard/compat/tensorflow_stub/dtypes.py:550: Futu
reWarning: Passing (type, 1) or '1type' as a synonym of type is depr
ecated; in a future version of numpy, it will be understood as (type
, (1,)) / '(1,)type'.
  np_resource = np.dtype([("resource", np.ubyte, 1)])
Using TensorFlow backend.
```

In [2]:
```
os.chdir("data/")
```

Helper methods for reading tweets and cleaning them.

In [3]:
```
def read_tsv(file_path):
    df = pd.read_table(file_path)
    return df

import string
import re

# code inspired from https://www.kaggle.com/rahulvv/bidirectional-lstm
-glove200d


def remove_urls(text):
    url = re.compile(r'https?://\S+|www\.\S+')
    return url.sub(r'',text)

def remove_html(text):
    html=re.compile(r'<.*?>')
    return html.sub(r'',text)

def split_text(text):
    text = text.split()
    return text

def lower(text):
    text = [word.lower() for word in text]
    return str(text)

def remove_punct(text):
    text = ''.join([char for char in text if char not in string.punctu
ation])
    text = re.sub('[0-9]+', '', str(text))
    return text

def remove_stopwords(text):
```

```
    pattern = re.compile(r'\b('+r'|'.join(stopwords.words('english'))
+ r')\b\s*')
    text = pattern.sub(' ', text)
    return text

lemmatizer = WordNetLemmatizer()
def lemmatize_words(text):
    text = lemmatizer.lemmatize(text)
    return text

def clean_tweet(text):
    t0 = remove_urls(text)
    t1 = remove_html(t0)
    t2 = split_text(t1)
    t3 = lower(t2)
    t4 = remove_punct(t3)
    t5 = remove_stopwords(t4)
    t6 = lemmatize_words(t5)
    return t6
```

In [4]:
```
tweet_df = pd.DataFrame(columns=['tweet', 'sentiment','NA'])
df_test = pd.DataFrame(columns=['tweet', 'sentiment','NA'])

for file in glob.glob("*.tsv"):
        if 'final_test' in file:
            df_test_cur = read_tsv(file)
            df_test = pd.concat([df_test, df_test_cur])
        else:
            df_train_cur = read_tsv(file)
            tweet_df = pd.concat([tweet_df, df_train_cur])
```

In [5]:
```
print(tweet_df[['tweet', 'sentiment']] )
```

```
                                                  tweet sentiment
0     05 Beat it - Michael Jackson - Thriller (25th ...   neutral
1     Jay Z joins Instagram with nostalgic tribute t...  positive
2     Michael Jackson: Bad 25th Anniversary Edition ...   neutral
3     I liked a @YouTube video http://t.co/AaR3pjp2P...  positive
4     18th anniv of Princess Diana's death. I still ...  positive
...                                                 ...       ...
1137                    Maybe it was - his - fantasy ?  positive
1138  It was ok , but they always just seem so nervo...  negative
1139  It is streamable from YepRoc -- matter of fact...  positive
1140  comment telling me who you are , or how you fo...  positive
1141  im on myspace ... ill try and find you and add...   neutral

[53368 rows x 2 columns]
```

In [6]:
```
print(df_test[['tweet', 'sentiment']] )
```

```
                                                   tweet sentiment
0      #ArianaGrande Ari By Ariana Grande 80% Full ht...   neutral
1      Ariana Grande KIIS FM Yours Truly CD listening...  positive
2      Ariana Grande White House Easter Egg Roll in W...  positive
3      #CD #Musics Ariana Grande Sweet Like Candy 3.4...  positive
4      SIDE TO SIDE 😘 @arianagrande #sidetoside #aria...   neutral
...                                                  ...       ...
11901  @dansen17 update: Zac Efron kissing a puppy ht...  positive
11902  #zac efron sex pic skins michelle sex https://...   neutral
11903  First Look at Neighbors 2 with Zac Efron Shirt...   neutral
11904  zac efron poses nude #lovely libra porn https:...   neutral
11905  #Fashion #Style The Paperboy (NEW Blu-ray Disc...   neutral

[11906 rows x 2 columns]
```

Reading Glove word embeddings into a dictionary.

In [7]:
```
#preparing train lables
tweet_df.loc[tweet_df.sentiment == "positive", "sentiment"] = 2
tweet_df.loc[tweet_df.sentiment == "neutral", "sentiment"] = 1
tweet_df.loc[tweet_df.sentiment == "negative", "sentiment"] = 0

labels = tweet_df["sentiment"].tolist()
labels = [ int(x) for x in labels ]

#preparing test labels
df_test.loc[df_test.sentiment == "positive", "sentiment"] = 2
df_test.loc[df_test.sentiment == "neutral", "sentiment"] = 1
df_test.loc[df_test.sentiment == "negative", "sentiment"] = 0

labels_test = df_test["sentiment"].tolist()
labels_test = [ int(x) for x in labels_test ]
```

Converting tweets and labels into lists.

In [8]:
```
train_tweets = tweet_df.tweet.values
y_train_orig = tweet_df.sentiment.values
test_tweets = df_test.tweet.values
```

In [9]:
```
from keras.utils import to_categorical

train_labels = to_categorical(y_train_orig)

clean_training_tweets = []
for i in range(len(train_tweets)):
    data = clean_tweet(train_tweets[i])
    clean_training_tweets.append(data)

clean_testing_tweets = []
for i in range(len(test_tweets)):
    data = clean_tweet(test_tweets[i])
    clean_testing_tweets.append(data)
```

Checking the tweets after cleaning them.

In [10]:
```
print(clean_training_tweets[:10])
print(clean_testing_tweets[:10])
```

```
[' beat  michael jackson  thriller th anniversary edition hd', 'jay
z joins instagram  nostalgic tribute  michael jackson jay z apparent
ly joined instagram  saturday  ', 'michael jackson bad th anniversar
y edition picture vinyl  unique picture disc vinyl includes  origina
l ', ' liked  youtube video one direction singing man  mirror  mich
ael jackson  atlanta ga june ', 'th anniv  princess dianas death  st
ill want  believe  living  private island away  public  michael j
ackson', 'oridaganjazz  st time  heard michael jackson sing  honolu
lu hawaii  restaurant  radio  abc  loved  ', 'michael jackson ap
peared  saturday  th place  top  miamis trends trndnl', '  old en
ough  remember michael jackson attending  grammys  brooke shields  w
ebster sat  lap  show', 'etbowser  u enjoy  nd rate michael jackso
n bit honest ques like  cant feel face song  god  obvious  want mj
', ' weeknd  closest thing  may get  michael jackson  long timeesp
ecially since  damn near mimics everything']
['arianagrande ari  ariana grande  full singer actress', 'ariana gra
nde kiis fm  truly cd listening party  burbank arianagrande', 'arian
a grande white house easter egg roll  washington arianagrande', 'cd
musics ariana grande sweet like candy  oz  ml sealed  box  authenic
new', 'side  side 😘 arianagrande sidetoside arianagrande musically
comunidadgay lgbt🌈 lotb…', 'hairspray live previews  macys thanksg
iving day parade arianagrande televisionnbc', 'lindsaylohan  ‘feelin
g thankful’  blasting arianagrande  wearing ‘toomuch…', ' hate  lo
ve  songs dammit arianagrande', 'ariana grande 【right  ft big sean】
アリアナ arianagrande', ' one would  prefer  listen  whole day 😍🤘
could never choose arianagrande intoyou sidetoside songs poll']
```

In [11]:
```
from nltk.probability import ConditionalFreqDist
from nltk.probability import FreqDist
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.linear_model import LogisticRegression
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import classification_report, confusion_matrix, a
ccuracy_score
```

In [15]:
```
all_tweets = clean_training_tweets + clean_testing_tweets

length = len(clean_training_tweets)
```

In [17]:
```
cv = CountVectorizer(binary=True, ngram_range = (1,3))
bow= cv.fit_transform(all_tweets)
bow_train = bow[:length]
bow_test = bow[length:]
```

In [22]:
```
model = MultinomialNB(alpha= 1.0).fit(bow_train, labels)
label_pred = model.predict(bow_test)

print("Classification Report for Naive Bayes")
print(confusion_matrix(labels_test,label_pred))
print(classification_report(labels_test,label_pred))
print(accuracy_score(labels_test, label_pred))
```

```
Classification Report for Naive Bayes
[[2831  692  288]
 [2412 2217 1114]
 [ 404  515 1433]]
              precision    recall  f1-score   support

           0       0.50      0.74      0.60      3811
           1       0.65      0.39      0.48      5743
           2       0.51      0.61      0.55      2352

    accuracy                           0.54     11906
   macro avg       0.55      0.58      0.54     11906
weighted avg       0.57      0.54      0.53     11906

0.5443473878716614
```