1 Tweet Analysis - Apple and Google

Author: Joseph Denney

Email: joseph.d.denney@gmail.com (mailto:joseph.d.denney@gmail.com)

github: www.github.com/josephdenney/Tweet_Analysis (http://www.github.com/josephdenney/Tweet_Analysis)

1.1 Introduction

1.1.1 Problem and Purpose

A client is looking to design and manufacture a new smart phone and will invariably compete with Apple and Google products. They have provided us with a data set of Tweets and would like more detail regarding negatively and positively charged Tweets directed at both iPhone OS and Android OS phones.

Our challenges are -

**** 1. To highlight any negative features of iPhones and Androids so that they can reduce them in their new product and***

**** 2. To highlight positive features of iPhones and Androids so that they can implement or improve them in their own product***

**** 3. To provide recommendations that will improve their future product***

1.2 Table of Contents

1.9 Question 1 and Recommendation

1.10 Question 2 and Recommendation

1.11 Question 3 and Recommendation

1.3 EDA and Data Preprocessing

1.3.1 Library, function, and data imports

In [44]:

```
import numpy as np
import pandas as pd
import spacy
import re
import nltk
import matplotlib.pyplot as plt
import logging
logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s',
                    level=logging.INFO)

from gensim.models import Word2Vec
from keras.models import Sequential
from keras.layers import Dense
from sklearn.preprocessing import MinMaxScaler, MaxAbsScaler
import seaborn as sns

from nltk.stem.wordnet import WordNetLemmatizer
import string
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('wordnet')
from sklearn.pipeline import Pipeline
from nltk.corpus import stopwords
from nltk import word_tokenize, FreqDist
from applesauce import model_scoring, cost_benefit_analysis, evaluate_model
from applesauce import model_opt, single_model_opt

from sklearn.metrics import classification_report, confusion_matrix
from sklearn.metrics import plot_confusion_matrix, accuracy_score
from sklearn.metrics import precision_recall_curve, f1_score, precision_score
from sklearn.metrics import recall_score
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.naive_bayes import BernoulliNB, CategoricalNB, GaussianNB
from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import TfidfVectorizer,CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.model_selection import GridSearchCV, train_test_split
from sklearn.utils import resample

from keras.preprocessing.sequence import pad_sequences
from keras.layers import Input, Dense, LSTM, Embedding
from keras.layers import Dropout, Activation, Bidirectional, GlobalMaxPool1D
from keras.models import Sequential
from keras import initializers, regularizers, constraints, optimizers, layers
from keras.preprocessing import text, sequence
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\josep\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\josep\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\josep\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

In [45]:
```
pattern = r"\@\w*"
```

In [46]:
```
# regex has a cool function called search
mysearch = re.search(pattern, df['Tweet'][0])
mysearch
```

Out[46]:
```
<re.Match object; span=(1, 10), match='@wesley83'>
```

In [47]:
```
myfind = re.findall(pattern, df['Tweet'][0])
myfind
```

Out[47]:
```
['@wesley83']
```

In [ ]:
```
```

In [48]:
```
nlp = spacy.load("en_core_web_sm")
```

In [49]:
```
print(stopwords)
print(nlp.Defaults.stop_words)
# view list of stopwords
```

```
<WordListCorpusReader in '.../corpora/stopwords' (not loaded yet)>
{'’d', 'since', 'thru', 'everywhere', 'within', 'forty', 'four', '‘ll', 'va
rious', 'all', 'own', 'same', 'your', 'this', 'two', 'perhaps', 'indeed',
'against', 'regarding', 'before', 'more', 'yours', 'made', 'an', 'again',
'am', 'by', 'beforehand', 'make', 'than', 'yourself', 'third', 'is', 'ofte
n', 'there', 'empty', 'you', 'another', 'have', "'m", 'n’t', 'out', 'someho
w', 'whence', 'either', 'after', 'anyone', 'back', 'due', 'onto', 'next',
'can', 'would', 'which', 'used', 'nine', 'whom', 'beyond', 'becomes', 'no',
'whose', 'unless', 'latterly', 'namely', 'while', 'whenever', 'in', 'betwee
n', 'done', 'last', 'least', 'up', 'over', 'say', 'were', 'below', 'alone',
'both', 'seeming', 'upon', 'wherein', '‘m', 'then', 'sixty', 'hereafter',
"'s", 'nowhere', 'first', 'themselves', 'something', 'and', 'yourselves',
'‘re', 'whether', 'do', 'off', 'full', 'did', 'many', '‘s', 'already', 'ho
w', 'seem', 'along', 'until', 'around', 'yet', 'ca', 'it', 'whereas', 'bein
g', 'take', 'what', 'their', "n't", 'ours', 'does', 'give', 'everyone', 'th
em', 'front', 'top', 'cannot', 'whoever', 'above', 'my', 'are', 'that', 'qu
ite', 'seemed', 'he', 'sometime', 'well', '‘ve', 'really', 'each', 'among',
'will', 'should', 'wherever', 'they', 'towards', '’ll', 'when', 'whereupo
n', 'still', 'afterwards', "'re", 'nobody', 'become', 'few', 'these', 'fift
y', 'elsewhere', 'per', '’re', 'further', 'must', 'be', 'myself', 'anythin
g', 'neither', 'mine', 'other', 'amount', 'n‘t', 'very', 'rather', 'mostl
y', 'anywhere', 'if', '’ve', 'meanwhile', 'ourselves', 'throughout', 'not',
'move', 'doing', 'too', 'bottom', 'even', 'who', 'otherwise', 'down', 'howe
ver', 'sometimes', 'therein', 'beside', 'anyhow', 'but', 'becoming', 'witho
ut', 'latter', 'except', 'part', 'fifteen', 'our', 'almost', 'the', 'now',
'call', '’m', 'became', 'for', 'get', 'thence', 'thereupon', 'him', 'moreov
er', 'hence', 'hereby', 'besides', 'eleven', 'hers', 'where', 'has', 'altho
ugh', 'noone', 'most', 'three', 'whereafter', 'serious', "'ve", 'eight', 'n
ame', 'every', 'of', 'nothing', 'into', 'amongst', 'somewhere', 'also', 'th
ose', 'twelve', 'please', 'one', 'ten', 'else', 'as', 'hereupon', 'whole',
'a', 'thus', 'her', 'behind', 'about', 'therefore', 'across', 'nor', 'other
s', 'us', 'much', 'using', 'several', 'i', 'former', 'go', 'herein', 'on',
'been', 'hundred', 'or', 'five', 'toward', 'his', 'only', 'to', 'always',
'such', 'herself', 'once', 'see', "'ll", 'here', 'was', 'she', 'just', 'nev
er', 'any', 'together', 'during', 'whither', 'so', 'whereby', 'neverthele
s', 'thereafter', 'at', 'with', 'though', 'had', 'anyway', 'none', 'six',
'from', 'someone', 'because', 'twenty', 'itself', "'d", 'show', 'thereby',
'me', 'whatever', 'under', 'we', 'why', 'ever', 'seems', 'some', 'through',
'keep', 'himself', '‘d', 'could', 'less', 'enough', 'put', 'may', 'everythi
ng', 'side', 're', 'its', 'might', 'via', '’s', 'formerly'}
```

In [50]:
```
df = pd.read_csv('data/product_tweets.csv',encoding='latin1')
```

In [51]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9093 entries, 0 to 9092
Data columns (total 3 columns):
 #   Column                                              Non-Null Count  Dt
ype
---  ------                                              --------------  --
---
 0   tweet_text                                          9092 non-null   ob
ject
 1   emotion_in_tweet_is_directed_at                     3291 non-null   ob
ject
 2   is_there_an_emotion_directed_at_a_brand_or_product  9093 non-null   ob
ject
dtypes: object(3)
memory usage: 213.2+ KB
```

In [52]:
```
df.head()
```

Out[52]:

	tweet_text	emotion_in_tweet_is_directed_at	is_there_an_emotion_directed_at_a_brand_or_pr
0	.@wesley83 I have a 3G iPhone. After 3 hrs twe...	iPhone	Negative er
1	@jessedee Know about @fludapp ? Awesome iPad/i...	iPad or iPhone App	Positive er
2	@swonderlin Can not wait for #iPad 2 also. The...	iPad	Positive er
3	@sxsw I hope this year's festival isn't as cra...	iPad or iPhone App	Negative er
4	@sxtxstate great stuff on Fri #SXSW: Marissa M...	Google	Positive er

In [53]:
```
df['emotion_in_tweet_is_directed_at'].unique()
```

Out[53]:
```
array(['iPhone', 'iPad or iPhone App', 'iPad', 'Google', nan, 'Android',
       'Apple', 'Android App', 'Other Google product or service',
       'Other Apple product or service'], dtype=object)
```

In [54]:
```
df['emotion_in_tweet_is_directed_at'].count()
```

Out[54]:
```
3291
```

1.3.2 Data Exploration and Column Title Cleanup

In [55]:
```
df['is_there_an_emotion_directed_at_a_brand_or_product'].unique()
```

Out[55]:
```
array(['Negative emotion', 'Positive emotion',
       'No emotion toward brand or product', "I can't tell"], dtype=object)
```

In [56]:
```
df = df.rename(columns= {'is_there_an_emotion_directed_at_a_brand_or_product'
                        :'Emotion',
                        'emotion_in_tweet_is_directed_at': 'Platform'})
```

In [57]:
```
df = df.rename(columns= {'tweet_text': 'Tweet'})
```

In [58]:
```
df.head()
```

Out[58]:

	Tweet	Platform	Emotion
0	.@wesley83 I have a 3G iPhone. After 3 hrs twe...	iPhone	Negative emotion
1	@jessedee Know about @fludapp ? Awesome iPad/i...	iPad or iPhone App	Positive emotion
2	@swonderlin Can not wait for #iPad 2 also. The...	iPad	Positive emotion
3	@sxsw I hope this year's festival isn't as cra...	iPad or iPhone App	Negative emotion
4	@sxtxstate great stuff on Fri #SXSW: Marissa M...	Google	Positive emotion

In [59]:
```
df.groupby(df['Platform']).count()
```

Out[59]:

	Tweet	Emotion
Platform		
Android	78	78
Android App	81	81
Apple	661	661
Google	430	430
Other Apple product or service	35	35
Other Google product or service	293	293
iPad	946	946
iPad or iPhone App	470	470
iPhone	297	297

1.3.3 Dummify Target Column

In [60]:
```
df_dummify = pd.get_dummies(df['Emotion'])
```

In [61]:
```
df_dummify.head()
```

Out[61]:

	I can't tell	Negative emotion	No emotion toward brand or product	Positive emotion
0	0	1	0	0
1	0	0	0	1
2	0	0	0	1
3	0	1	0	0
4	0	0	0	1

In [62]:
```
df_dummify.sum() # class bias
```

Out[62]:
```
I can't tell                           156
Negative emotion                       570
No emotion toward brand or product    5389
Positive emotion                      2978
dtype: int64
```

In [63]:
```
df.info()
df = pd.merge(df, df_dummify, how='outer',on=df.index)
# ran this code, dummify emotion data
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9093 entries, 0 to 9092
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Tweet     9092 non-null   object
 1   Platform  3291 non-null   object
 2   Emotion   9093 non-null   object
dtypes: object(3)
memory usage: 213.2+ KB
```

In [64]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9093 entries, 0 to 9092
Data columns (total 8 columns):
 #   Column                              Non-Null Count  Dtype 
---  ------                              --------------  ----- 
 0   key_0                               9093 non-null   int64 
 1   Tweet                               9092 non-null   object
 2   Platform                            3291 non-null   object
 3   Emotion                             9093 non-null   object
 4   I can't tell                        9093 non-null   uint8 
 5   Negative emotion                    9093 non-null   uint8 
 6   No emotion toward brand or product  9093 non-null   uint8 
 7   Positive emotion                    9093 non-null   uint8 
dtypes: int64(1), object(3), uint8(4)
memory usage: 390.7+ KB
```

In [65]:
```
df.head()
```

Out[65]:

	key_0	Tweet	Platform	Emotion	I can't tell	Negative emotion	No emotion toward brand or product	Positive emotion
0	0	.@wesley83 I have a 3G iPhone. After 3 hrs twe...	iPhone	Negative emotion	0	1	0	0
1	1	@jessedee Know about @fludapp ? Awesome iPad/i...	iPad or iPhone App	Positive emotion	0	0	0	1
2	2	@swonderlin Can not wait for #iPad 2 also. The...	iPad	Positive emotion	0	0	0	1
3	3	@sxsw I hope this year's festival isn't as cra...	iPad or iPhone App	Negative emotion	0	1	0	0
4	4	@sxtxstate great stuff on Fri #SXSW: Marissa M...	Google	Positive emotion	0	0	0	1

In [66]:
```
df = df.rename(columns = {"I can't tell": "Uncertain",
                          'Negative emotion': 'Negative',
                          'No emotion toward brand or product': 'No Emotion',
                          'Positive emotion':'Positive'})
```

In [67]:
```
df = df.drop(columns='key_0')
df.head()
df.to_csv('Full_DF')
```

In [68]:
```
corpus = list(df['Tweet']) # verify corpus list
corpus[:10]
```

Out[68]:
```
['.@wesley83 I have a 3G iPhone. After 3 hrs tweeting at #RISE_Austin, it was dead!  I need to upgrade. Plugin stations at #SXSW.',
 "@jessedee Know about @fludapp ? Awesome iPad/iPhone app that you'll likely appreciate for its design. Also, they're giving free Ts at #SXSW",
 '@swonderlin Can not wait for #iPad 2 also. They should sale them down at #SXSW.',
 "@sxsw I hope this year's festival isn't as crashy as this year's iPhone app. #sxsw",
 "@sxtxstate great stuff on Fri #SXSW: Marissa Mayer (Google), Tim O'Reilly (tech books/conferences) &amp; Matt Mullenweg (Wordpress)",
 '@teachntech00 New iPad Apps For #SpeechTherapy And Communication Are Showcased At The #SXSW Conference http://ht.ly/49n4M (http://ht.ly/49n4M) #iear #edchat #asd',
 nan,
 '#SXSW is just starting, #CTIA is around the corner and #googleio is only a hop skip and a jump from there, good time to be an #android fan',
 'Beautifully smart and simple idea RT @madebymany @thenextweb wrote about our #hollergram iPad app for #sxsw! http://bit.ly/ieaVOB', (http://bit.ly/ieaVOB',)
 'Counting down the days to #sxsw plus strong Canadian dollar means stock up on Apple gear']
```

1.3.4 Platform Negative Tweet Table

In [69]:
```
df.groupby(by=df['Platform']).sum()
```

Out[69]:

Platform	Uncertain	Negative	No Emotion	Positive
Android	0.0	8.0	1.0	69.0
Android App	0.0	8.0	1.0	72.0
Apple	2.0	95.0	21.0	543.0
Google	1.0	68.0	15.0	346.0
Other Apple product or service	0.0	2.0	1.0	32.0
Other Google product or service	1.0	47.0	9.0	236.0
iPad	4.0	125.0	24.0	793.0
iPad or iPhone App	0.0	63.0	10.0	397.0
iPhone	1.0	103.0	9.0	184.0

1.3.5 Tokenize and Create Bag of Words

In [70]:
```
tokenz = word_tokenize(','.join(str(v) for v in corpus))
```

In [71]:
```
tokenz[:10]
```
Out[71]:
```
['.', '@', 'wesley83', 'I', 'have', 'a', '3G', 'iPhone', '.', 'After']
```

1.3.6 Create Stopwords List

In [72]:
```
stopword_list = list(nlp.Defaults.stop_words)
len(nlp.Defaults.stop_words)
```
Out[72]:
```
326
```

In [73]:
```
stopword_list
```
Out[73]:
```
['’d',
 'since',
 'thru',
 'everywhere',
 'within',
 'forty',
 'four',
 '‘ll',
 'various',
 'all',
 'own',
 'same',
 'your',
 'this',
 'two',
 'perhaps',
 'indeed',
 'against',
 'regarding',
```

In [74]:
```
stopword_list.extend(string.punctuation)
```

In [75]:
```
len(stopword_list)
```
Out[75]:
```
358
```

In [76]:
```
stopword_list.extend(stopwords.words('english'))
```

In [77]:
```
len(stopword_list)
```
Out[77]:
```
537
```

In [78]:
```
additional_punc = ['“','”','...',"''",'’','``','https','rt','\.+']
stopword_list.extend(additional_punc)
stopword_list[-10:]
```
Out[78]:
```
["wouldn't", '“', '”', '...', "''", '’', '``', 'https', 'rt', '\\.+']
```

1.3.7 Remove Stopwords and Additional Punctuation from the

Data

In [79]:
```
stopped_tokenz = [word.lower() for word in tokenz if word.lower()
                  not in stopword_list]
```

In [80]:
```
freq = FreqDist(stopped_tokenz)
freq.most_common(50)
```

Out[80]:
```
[('sxsw', 9418),
 ('mention', 7120),
 ('link', 4313),
 ('google', 2593),
 ('ipad', 2432),
 ('apple', 2301),
 ('quot', 1696),
 ('iphone', 1516),
 ('store', 1472),
 ('2', 1114),
 ('new', 1090),
 ('austin', 959),
 ('amp', 836),
 ('app', 810),
 ('circles', 658),
 ('launch', 653),
 ('social', 647),
 ('android', 574),
 ('today', 574),
 ('network', 465),
 ('ipad2', 457),
 ('pop-up', 420),
 ('line', 405),
 ('free', 387),
 ('called', 361),
 ('party', 346),
 ('sxswi', 340),
 ('mobile', 338),
 ('major', 301),
 ('like', 290),
 ('time', 271),
 ('temporary', 264),
 ('opening', 257),
 ('possibly', 240),
 ('people', 226),
 ('downtown', 225),
 ('apps', 224),
 ('great', 222),
 ('maps', 219),
 ('going', 217),
 ('check', 216),
 ('mayer', 214),
 ('day', 214),
 ('open', 210),
 ('popup', 209),
 ('need', 205),
 ('marissa', 189),
 ('got', 185),
 ('w/', 182),
 ('know', 180)]
```

1.3.8 Lemmatize the Data, Utilize Regex to Find and Remove URL's, Tags, other Misc

URL's, Tags, other Misc

In [81]:
```
additional_misc = ['sxsw','mention',r'[a-zA-Z]+\'?s]',
                   r"(http[s]?://\w*\.\w*/+\w+)", r'\#\w*',
                   r'RT [@]?\w*:', r'\@\w*',r"\d$",r"^\d",
                   r"([a-zA-Z]+(?:'[a-z]+)?)",r'\d.',r'\d','RT',
                   r'^http[s]?','za'] #[A-Z]{2,20} remove caps like MAGA and
stopword_list.extend(additional_misc)
stopword_list.extend(['0', '1', '2', '3', '4', '5', '6', '7', '8', '9'])
```

In [82]:
```
additional_misc = [r'\@\w*']
stopword_list.extend(additional_misc)
stopword_list.extend(['0', '1', '2', '3', '4', '5', '6', '7', '8', '9'])
```

In [ ]:
```

```

In [83]:
```
lemmatizer = WordNetLemmatizer()
```

In [84]:
```
clean_stopped_tokenz = [word.lower() for word in stopped_tokenz if word
                        not in stopword_list]
clean_lemmatized_tokenz = [lemmatizer.lemmatize(word.lower()) for word
                           in stopped_tokenz if word not in stopword_list]
```

In [85]:
```
freq_clean_lemma = FreqDist(clean_lemmatized_tokenz)
freq_lemma = freq_clean_lemma.most_common(5000)
freq_lemma2 = freq_clean_lemma.most_common(25)
```

In [86]:
```
total_word_count = len(clean_lemmatized_tokenz)
```

In [87]:
```
lemma_word_count = sum(freq_clean_lemma.values()) # just a number
```

In [88]:

```
for word in freq_lemma2: # separate both classes, positive and negative
    normalized_freq = word[1] / lemma_word_count
    print(word, "----", "{:.3f}".format(normalized_freq*100),"%")
```

```
('link', 4324) ---- 5.004 %
('google', 2594) ---- 3.002 %
('ipad', 2432) ---- 2.814 %
('apple', 2304) ---- 2.666 %
('quot', 1696) ---- 1.963 %
('iphone', 1516) ---- 1.754 %
('store', 1511) ---- 1.749 %
('new', 1090) ---- 1.261 %
('austin', 960) ---- 1.111 %
('amp', 836) ---- 0.967 %
('app', 810) ---- 0.937 %
('launch', 691) ---- 0.800 %
('circle', 673) ---- 0.779 %
('social', 647) ---- 0.749 %
('android', 574) ---- 0.664 %
('today', 574) ---- 0.664 %
('network', 473) ---- 0.547 %
('ipad2', 457) ---- 0.529 %
('line', 442) ---- 0.512 %
('pop-up', 422) ---- 0.488 %
('free', 387) ---- 0.448 %
('party', 386) ---- 0.447 %
('called', 361) ---- 0.418 %
('mobile', 340) ---- 0.393 %
('sxswi', 340) ---- 0.393 %
```

In [89]:

```
# from wordcloud import WordCloud

# ## Initalize a WordCloud with our stopwords_list and no bigrams
# wordcloud = WordCloud(stopwords=stopword_list,collocations=False)

# ## Generate wordcloud from stopped_tokens
# wordcloud.generate(','.join(clean_lemmatized_tokenz))

# ## Plot with matplotlib
# plt.figure(figsize = (12, 12), facecolor = None)
# plt.imshow(wordcloud)
# plt.axis('off')
```

In [90]:

```
bigram_measures = nltk.collocations.BigramAssocMeasures()
tweet_finder = nltk.BigramCollocationFinder.from_words(clean_lemmatized_toker
tweets_scored = tweet_finder.score_ngrams(bigram_measures.raw_freq)
```

In [91]:

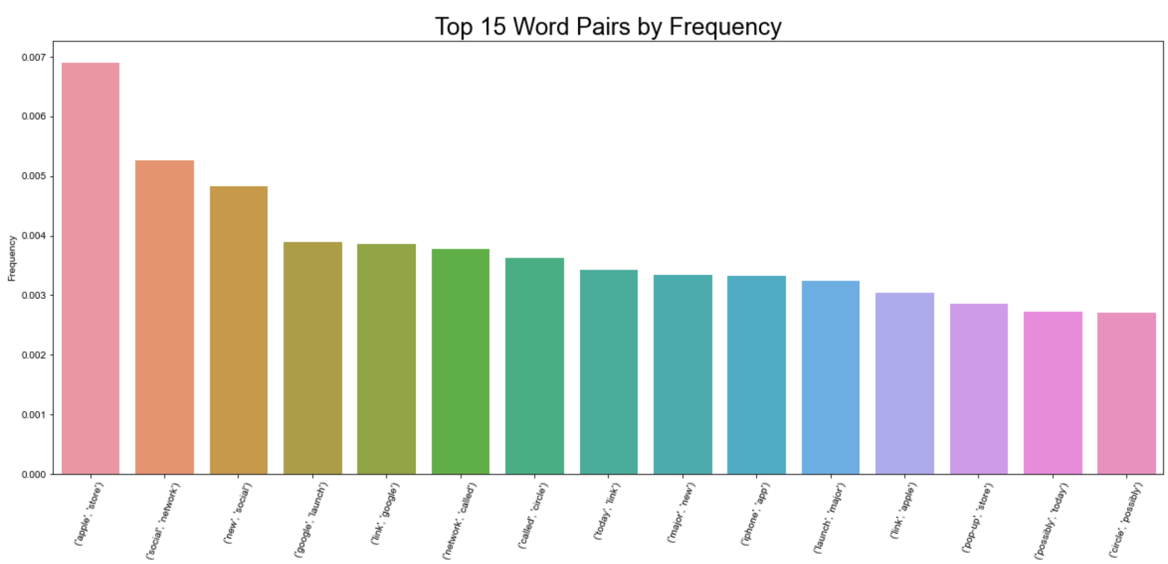
```
word_pairs = pd.DataFrame(tweets_scored, columns=["Word","Freq"]).head(20)
word_pairs
```

Out[91]:

	Word	Freq
0	(apple, store)	0.006920
1	(social, network)	0.005277
2	(new, social)	0.004837
3	(google, launch)	0.003912
4	(link, google)	0.003877
5	(network, called)	0.003784
6	(called, circle)	0.003634
7	(today, link)	0.003437
8	(major, new)	0.003356
9	(iphone, app)	0.003333
10	(launch, major)	0.003264

In [92]:

```
fig_dims = (20,8)
fig, ax = plt.subplots(figsize=fig_dims)
sns.set(font_scale=2)
sns.set_style("darkgrid")
palette = sns.set_palette("dark")
ax = sns.barplot(x=word_pairs.head(15)['Word'], y=word_pairs.head(15)['Freq']
                 palette=palette)
ax.set(xlabel="Word Pairs",ylabel="Frequency")
plt.ticklabel_format(style='plain',axis='y')
plt.xticks(rotation=70)
plt.title('Top 15 Word Pairs by Frequency')
plt.show()
```



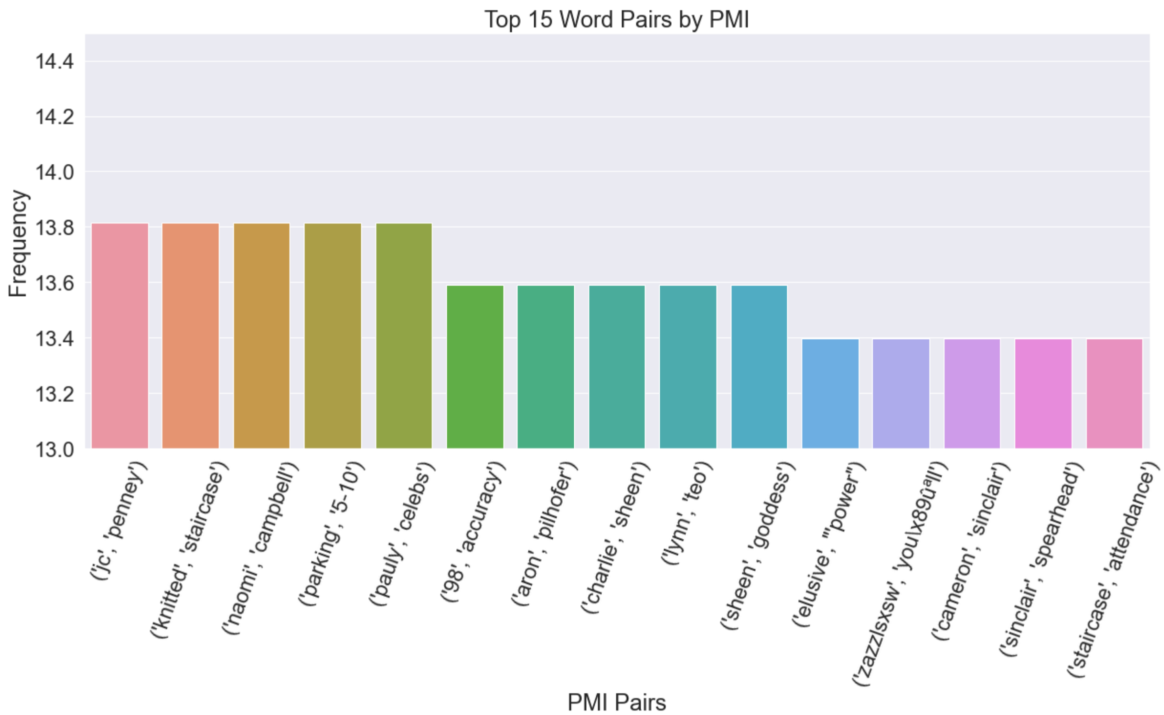
In [93]:
```
tweet_pmi_finder = nltk.BigramCollocationFinder.from_words(clean_lemmatized_t
tweet_pmi_finder.apply_freq_filter(5)

tweet_pmi_scored = tweet_pmi_finder.score_ngrams(bigram_measures.pmi)
```

In [94]:
```
PMI_list = pd.DataFrame(tweet_pmi_scored, columns=["Words","PMI"]).head(20)
PMI_list = PMI_list[PMI_list.PMI < 14]
PMI_list
```

Out[94]:

	Words	PMI
1	(jc, penney)	13.813948
2	(knitted, staircase)	13.813948
3	(naomi, campbell)	13.813948
4	(parking, 5-10)	13.813948
5	(pauly, celebs)	13.813948
6	(98, accuracy)	13.591556
7	(aron, pilhofer)	13.591556
8	(charlie, sheen)	13.591556
9	(lynn, teo)	13.591556
10	(sheen, goddess)	13.591556
11	(elusive, 'power)	13.398911

In [95]:
```
fig_dims = (20,8)
fig, ax = plt.subplots(figsize=fig_dims)
sns.set(font_scale=2)
sns.set_style("darkgrid")
palette = sns.set_palette("dark")
ax = sns.barplot(x=PMI_list.head(15)['Words'], y=PMI_list.head(15)['PMI'],
                 palette=palette)
ax.set(xlabel="PMI Pairs",ylabel="Frequency")
plt.ylim([13,14.5])
plt.ticklabel_format(style='plain',axis='y')
plt.xticks(rotation=70)
plt.title('Top 15 Word Pairs by PMI')
plt.show()
```



In [96]:
```
df1 = df
df.head()
```

Out[96]:

	Tweet	Platform	Emotion	Uncertain	Negative	No Emotion	Positive
0	.@wesley83 I have a 3G iPhone. After 3 hrs twe...	iPhone	Negative emotion	0	1	0	0
1	@jessedee Know about @fludapp ? Awesome iPad/i...	iPad or iPhone App	Positive emotion	0	0	0	1
2	@swonderlin Can not wait for #iPad 2 also. The...	iPad	Positive emotion	0	0	0	1
3	@sxsw I hope this year's festival isn't as cra...	iPad or iPhone App	Negative emotion	0	1	0	0
4	@sxtxstate great stuff on Fri #SXSW: Marissa M...	Google	Positive emotion	0	0	0	1

In [97]:
```
df1 = df1.drop(columns=['Uncertain','No Emotion'])
# Turn negative and positive columns into one column of just negatives
# and positive.
df1 = df1[df1['Emotion'] != "No emotion toward brand or product"]
df1 = df1[df1['Emotion'] != "I can't tell"]
df1 = df1.drop(columns='Negative')
df1 = df1.rename(columns={'Positive': 'Positive_Bin'})
df1.head()
```

Out[97]:

	Tweet	Platform	Emotion	Positive_Bin
0	.@wesley83 I have a 3G iPhone. After 3 hrs twe...	iPhone	Negative emotion	0
1	@jessedee Know about @fludapp ? Awesome iPad/i...	iPad or iPhone App	Positive emotion	1
2	@swonderlin Can not wait for #iPad 2 also. The...	iPad	Positive emotion	1
3	@sxsw I hope this year's festival isn't as cra...	iPad or iPhone App	Negative emotion	0
4	@sxtxstate great stuff on Fri #SXSW: Marissa M...	Google	Positive emotion	1

In [98]:
```
df1.to_csv('Tweet.csv')
```

1.3.9 Create Upsampled Data

In [99]:
```
df_majority = df1.loc[df1['Positive_Bin']==1]
df_minority = df1.loc[df1['Positive_Bin']==0]
```

In [100]:
```
df_minority.shape
```

Out[100]: (570, 4)

In [101]:
```
df_majority.shape
```

Out[101]: (2978, 4)

In [102]:
```
df_min_sample = resample(df_minority, replace=True, n_samples=1000,
                         random_state=42)
```

In [103]:
```
df_maj_sample = resample(df_majority, replace=True, n_samples=2500,
                         random_state=42)
```

In [104]:
```
df_upsampled = pd.concat([df_min_sample, df_maj_sample], axis=0)
df_upsampled.shape
```

Out[104]: (3500, 4)

In [105]:
```
X, y = df_upsampled['Tweet'], df_upsampled['Positive_Bin']
```

In [106]:
```
df_upsampled.to_csv('Upsampled.csv')
```

1.4 Modeling

1.4.1 Train/Test Split

In [107]:
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42)
```

In [108]:
```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3548 entries, 0 to 9088
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   Tweet         3548 non-null   object
 1   Platform      3191 non-null   object
 2   Emotion       3548 non-null   object
 3   Positive_Bin  3548 non-null   uint8 
dtypes: object(3), uint8(1)
memory usage: 114.3+ KB
```

In [109]:
```
y_train.value_counts(0)
y_test.value_counts(1)
```

```
2020-12-30 10:37:01,139 : INFO : NumExpr defaulting to 8 threads.
```

Out[109]:
```
1    0.683429
0    0.316571
Name: Positive_Bin, dtype: float64
```

1.4.2 Vectorize and Tokenize with Count Vectorizer and Tf Idf

In [110]:
```
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer,
from sklearn.ensemble import RandomForestClassifier

tokenizer = nltk.TweetTokenizer(preserve_case=False)

vectorizer = CountVectorizer(tokenizer=tokenizer.tokenize,
                             stop_words=stopword_list,decode_error='ignore')
```

In [111]:
```
X_train_count = vectorizer.fit_transform(X_train)
X_test_count = vectorizer.transform(X_test)
```

```
C:\Users\josep\anaconda3\lib\site-packages\sklearn\feature_extraction\text.
py:383: UserWarning: Your stop_words may be inconsistent with your preproce
ssing. Tokenizing the stop words generated tokens [":'[", ':/', 'a-z', 'a-z
a-z', 'http', 'n', 'w', '‘'] not in stop_words.
  warnings.warn('Your stop_words may be inconsistent with '
```

1.4.3 MaxAbsScaler

In [112]:
```
scaler_object = MaxAbsScaler().fit(X_train_count)
```

In [113]:
```
scaler_object.transform(X_train_count)
```

Out[113]:
```
<2625x4295 sparse matrix of type '<class 'numpy.float64'>'
        with 28229 stored elements in Compressed Sparse Row format>
```

In [114]:
```
scaler_object.transform(X_test_count)
```

Out[114]:
```
<875x4295 sparse matrix of type '<class 'numpy.float64'>'
        with 8854 stored elements in Compressed Sparse Row format>
```

1.4.4 Instantiate Model

In [115]:
```
ran_for = RandomForestClassifier(class_weight='balanced')
model = ran_for.fit(X_train_count, y_train)
```

In [116]:
```
y_hat_test = model.predict(X_test_count)
```

1.5 Evaluate Models

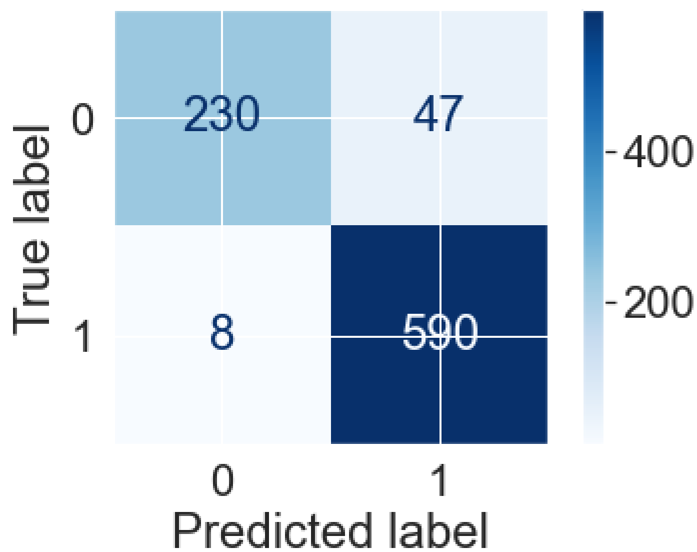
1 denotes a Positive Tweet, 0 denotes a Negative Tweet

1.5.1 Random Forest with Count Vectorizer

In [117]:

```
evaluate_model(y_test, y_hat_test, X_test_count,clf=model)
# 1 denotes Positive Tweet
```

```
              precision    recall  f1-score   support

           0       0.97      0.83      0.89       277
           1       0.93      0.99      0.96       598

    accuracy                           0.94       875
   macro avg       0.95      0.91      0.92       875
weighted avg       0.94      0.94      0.94       875
```



Basic Random Forest model performs well after preprocessing with high precision and f1-scores.

In [118]:

```
tf_idf_vectorizer = TfidfVectorizer(tokenizer=tokenizer.tokenize,
                                    stop_words=stopword_list,
                                    decode_error='ignore')
```

In [119]:
```python
X_train_tf_idf = tf_idf_vectorizer.fit_transform(X_train)
X_test_tf_idf = tf_idf_vectorizer.transform(X_test)
print(X_train_tf_idf.shape)
print(y_train.shape)
```

```
C:\Users\josep\anaconda3\lib\site-packages\sklearn\feature_extraction\text.
py:383: UserWarning: Your stop_words may be inconsistent with your preproce
ssing. Tokenizing the stop words generated tokens [":'[", ':/', 'a-z', 'a-z
a-z', 'http', 'n', 'w', '‘'] not in stop_words.
  warnings.warn('Your stop_words may be inconsistent with '

(2625, 4295)
(2625,)
```

In [120]:
```python
from sklearn.ensemble import RandomForestClassifier
```

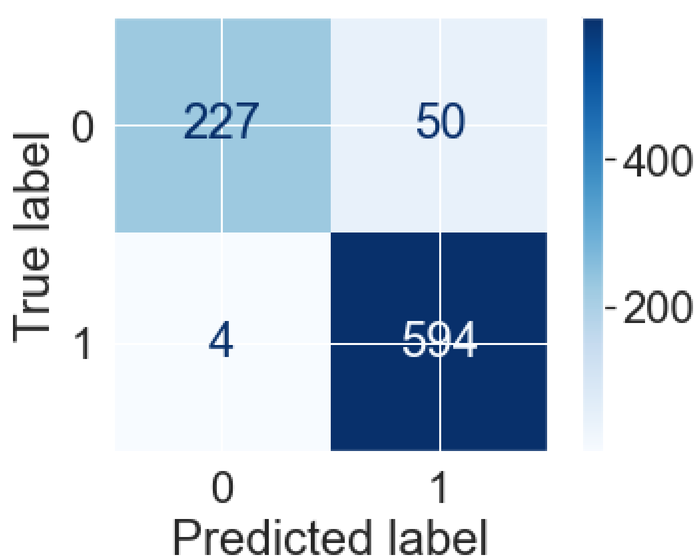
In [121]:
```python
ran_for = RandomForestClassifier(class_weight='balanced')
model_tf_idf = ran_for.fit(X_train_tf_idf,y_train)
```

In [122]:
```python
y_hat_tf_idf = model_tf_idf.predict(X_test_count)
```

1.5.2 Random Forest with Tf-Idf Vectorizer

In [123]:
```python
evaluate_model(y_test, y_hat_tf_idf, X_test_tf_idf,clf=model_tf_idf)
```

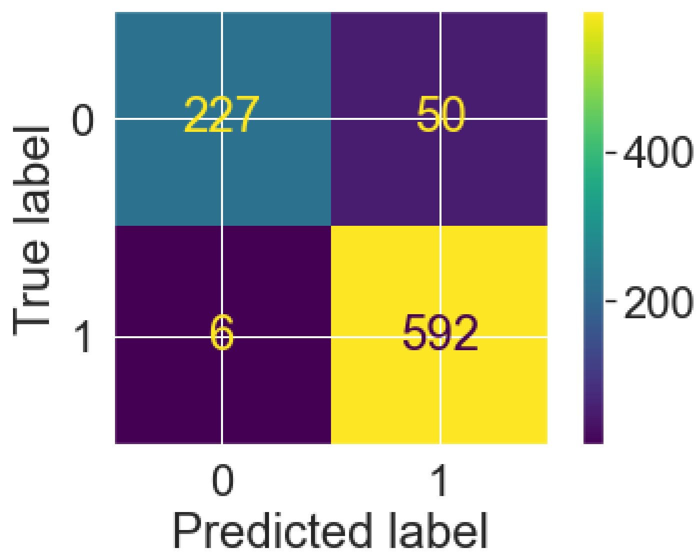
```
              precision    recall  f1-score   support

           0       0.91      0.61      0.73       277
           1       0.84      0.97      0.90       598

    accuracy                           0.86       875
   macro avg       0.88      0.79      0.82       875
weighted avg       0.87      0.86      0.85       875
```



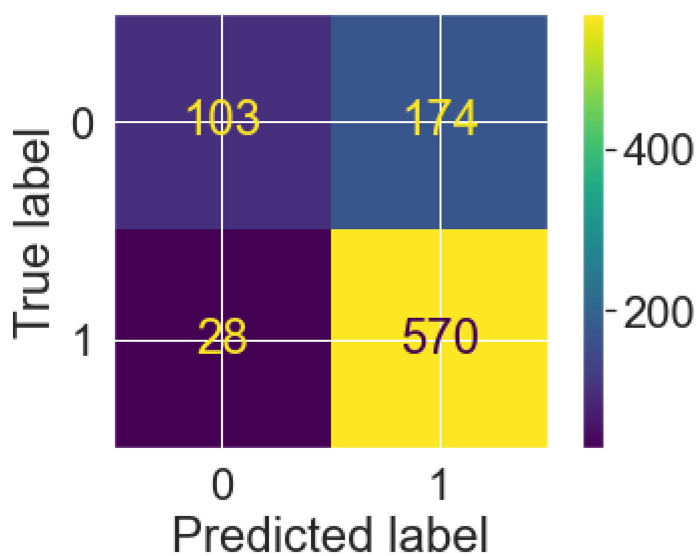
1.5.3 Multiple Models, CountVectorizer

In [124]:
```
ran_for = RandomForestClassifier()
ada_clf = AdaBoostClassifier()
gb_clf = GradientBoostingClassifier()

models = [ran_for, ada_clf, gb_clf]

for model in models:
    single_model_opt(model, X_train_count, y_train, X_test_count, y_test)
```

```
Accuracy Score:  0.936
Precision Score:  0.9221183800623053
Recall Score:  0.9899665551839465
F1 Score:  0.9548387096774194
RandomForestClassifier()   0.936
```

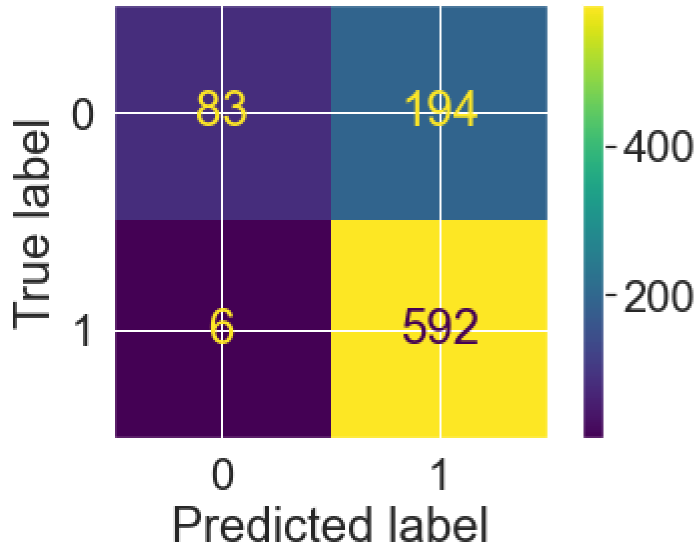



```
Accuracy Score:  0.7691428571428571
Precision Score:  0.7661290322580645
Recall Score:  0.9531772575250836
F1 Score:  0.849478390461997
AdaBoostClassifier()   0.7691428571428571
```




```
Accuracy Score:  0.7714285714285715
```

```
Precision Score:  0.7531806615776081
Recall Score:  0.9899665551839465
F1 Score:  0.8554913294797688
GradientBoostingClassifier()   0.7714285714285715
```

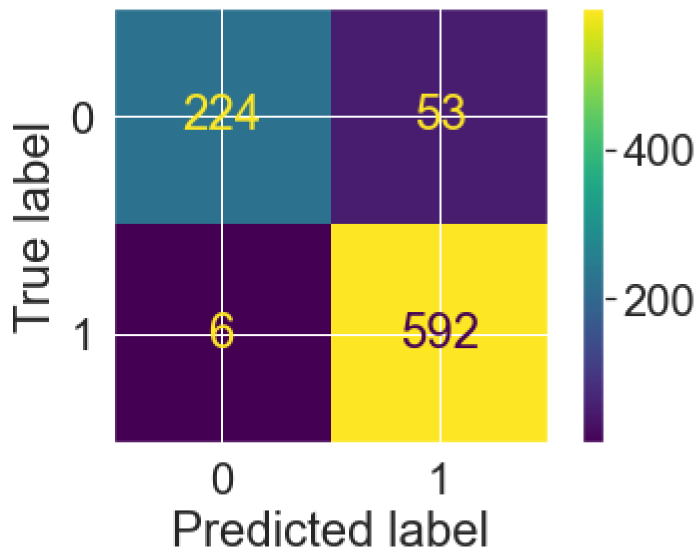


1.5.4 Multiple Models, Tf-Idf Vectorizer

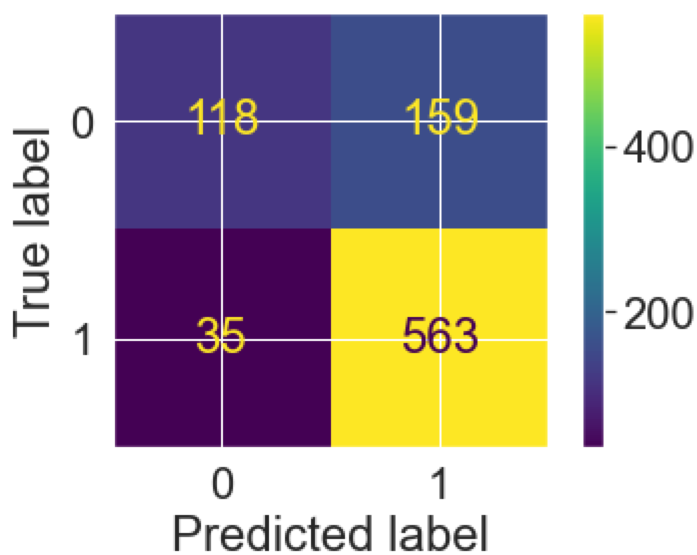
In [125]:

```
for model in models:
    single_model_opt(model, X_train_tf_idf, y_train, X_test_tf_idf, y_test)
```

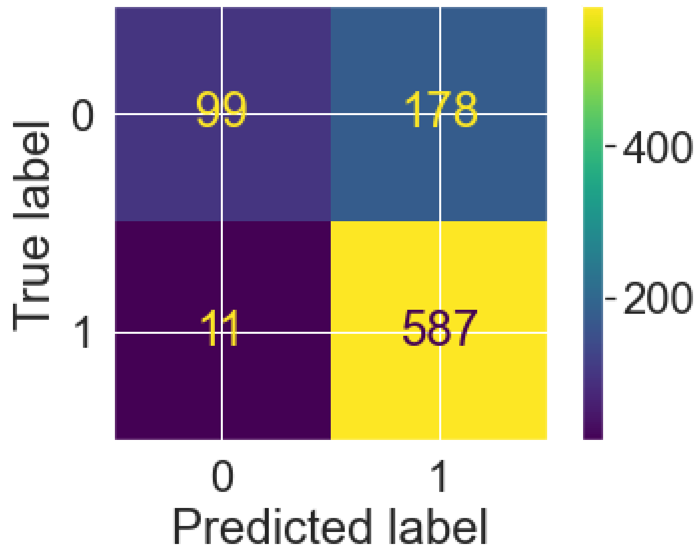
```
Accuracy Score:  0.9325714285714286
Precision Score:  0.9178294573643411
Recall Score:  0.9899665551839465
F1 Score:  0.9525341914722445
RandomForestClassifier()   0.9325714285714286
```




```
Accuracy Score:  0.7782857142857142
Precision Score:  0.7797783933518005
Recall Score:  0.9414715719063546
F1 Score:  0.8530303030303031
AdaBoostClassifier()   0.7782857142857142
```




```
Accuracy Score:  0.784
Precision Score:  0.7673202614379085
Recall Score:  0.9816053511705686
F1 Score:  0.861335289801907 7
GradientBoostingClassifier()   0.784
```

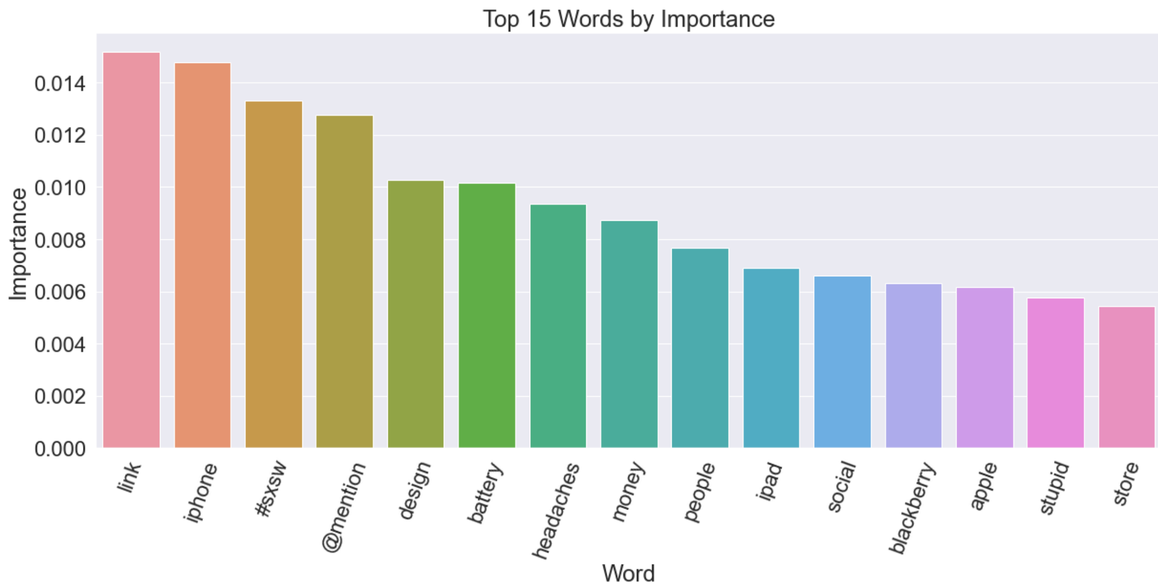


In [126]:
```
tf_idf_vectorizer.get_feature_names()
```

Out[126]:
```
['##sxsw',
 '#10',
 '#106',
 '#11ntc',
 '#1406-08',
 '#15slides',
 '#310409h2011',
 '#4sq',
 '#911tweets',
 '#abacus',
 '#accesssxsw',
 '#accordion',
 '#aclu',
 '#adam',
 '#addictedtotheinterwebs',
 '#adpeopleproblems',
 '#agchat',
 '#agileagency',
 '#agnerd',
```

In [127]:
```
importance = pd.Series(ran_for.feature_importances_,
                       index=tf_idf_vectorizer.get_feature_names())
importance = pd.DataFrame(importance).sort_values(by=0,ascending=False)
```

In [128]:
```
fig_dims = (20,8)
fig, ax = plt.subplots(figsize=fig_dims)
sns.set(font_scale=2)
sns.set_style("darkgrid")
palette = sns.set_palette("dark")
ax = sns.barplot(x=importance.head(15).index, y=importance.head(15)[0],
                 palette=palette)
ax.set(xlabel="Word",ylabel="Importance")
plt.ticklabel_format(style='plain',axis='y')
plt.xticks(rotation=70)
plt.title('Top 15 Words by Importance')
plt.show()
```



1.5.5 Pipeline and GridSearchCV

In [129]:
```
vectorizer = CountVectorizer()
tf_transform = TfidfTransformer(use_idf=True)
```

In [130]:
```
text_pipe = Pipeline(steps=[
    ('count_vectorizer',vectorizer),
    ('tf_transformer',tf_transform)])
```

In [131]:
```
RandomForestClassifier(class_weight='balanced')
```

Out[131]:
```
RandomForestClassifier(class_weight='balanced')
```

In [132]:
```
full_pipe = Pipeline(steps=[
    ('text_pipe',text_pipe),
    ('clf',RandomForestClassifier(class_weight='balanced'))
])
```

In [133]:
```
X_train_pipe = text_pipe.fit_transform(X_train)
```

In [134]:
```
X_test_pipe = text_pipe.transform(X_test)
```

In [135]:
```
X_train_pipe
```

Out[135]:
```
<2625x4256 sparse matrix of type '<class 'numpy.float64'>'
	with 44273 stored elements in Compressed Sparse Row format>
```

In [136]:
```
params = {'text_pipe__tf_transformer__use_idf':[True, False],
          'text_pipe__count_vectorizer__tokenizer':[None,tokenizer.tokenize],
          'text_pipe__count_vectorizer__stop_words':[None,stopword_list],
          'clf__criterion':['gini', 'entropy']}
```

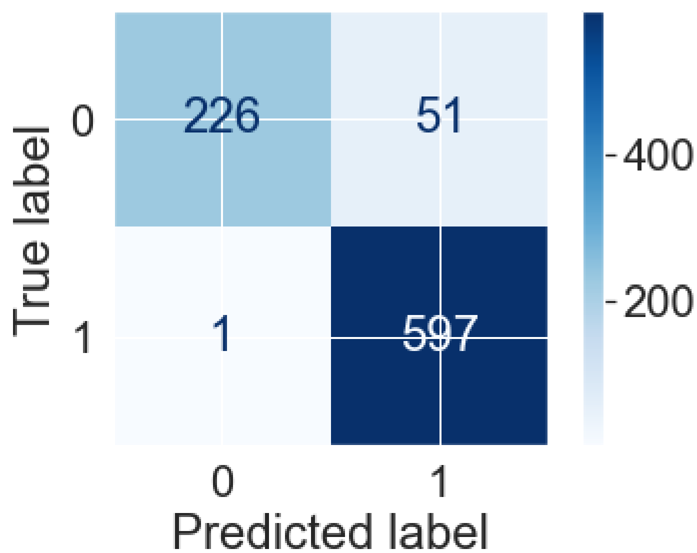
In [137]:
```
## Make and fit grid
grid = GridSearchCV(full_pipe,params,cv=3)
grid.fit(X_train,y_train)
## Display best params
grid.best_params_
```

```
C:\Users\josep\anaconda3\lib\site-packages\sklearn\feature_extraction\tex
t.py:383: UserWarning: Your stop_words may be inconsistent with your prep
rocessing. Tokenizing the stop words generated tokens ['http'] not in sto
p_words.
  warnings.warn('Your stop_words may be inconsistent with '
C:\Users\josep\anaconda3\lib\site-packages\sklearn\feature_extraction\tex
t.py:383: UserWarning: Your stop_words may be inconsistent with your prep
rocessing. Tokenizing the stop words generated tokens ['http'] not in sto
p_words.
  warnings.warn('Your stop_words may be inconsistent with '
C:\Users\josep\anaconda3\lib\site-packages\sklearn\feature_extraction\tex
t.py:383: UserWarning: Your stop_words may be inconsistent with your prep
rocessing. Tokenizing the stop words generated tokens ['http'] not in sto
p_words.
  warnings.warn('Your stop_words may be inconsistent with '
C:\Users\josep\anaconda3\lib\site-packages\sklearn\feature_extraction\tex
t.py:383: UserWarning: Your stop_words may be inconsistent with your prep
rocessing. Tokenizing the stop words generated tokens ['http'] not in sto
p_words.
```

In [138]:
```
best_pipe = grid.best_estimator_
y_hat_test = grid.predict(X_test)
```

In [139]:
```
evaluate_model(y_test,y_hat_test,X_test,best_pipe)
```

```
              precision    recall  f1-score   support

           0       1.00      0.82      0.90       277
           1       0.92      1.00      0.96       598

    accuracy                           0.94       875
   macro avg       0.96      0.91      0.93       875
weighted avg       0.94      0.94      0.94       875
```



In [140]:
```
X_train_pipe.shape
```

Out[140]: (2625, 4256)

1.5.6 Bigram Frequency

In [141]:
```
features = text_pipe.named_steps['count_vectorizer'].get_feature_names()
features[:10]
```

Out[141]: ['000', '02', '03', '0310apple', '08', '10', '100', '100s', '101', '106']

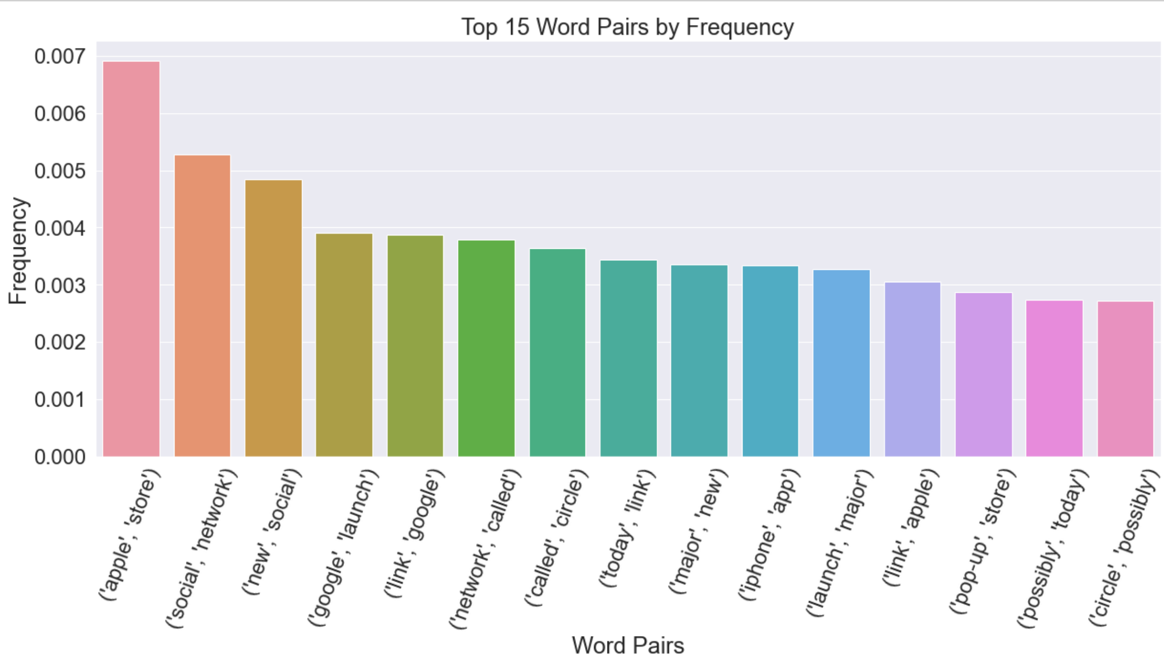
In [142]:
```
bigram_measures = nltk.collocations.BigramAssocMeasures()
tweet_finder = nltk.BigramCollocationFinder.from_words(clean_lemmatized_toker
tweets_scored = tweet_finder.score_ngrams(bigram_measures.raw_freq)
```

In [143]:
```
bigram1 = pd.DataFrame(tweets_scored, columns=['Words','Freq'])
bigram1.head()
```

Out[143]:

	Words	Freq
0	(apple, store)	0.006920
1	(social, network)	0.005277
2	(new, social)	0.004837
3	(google, launch)	0.003912
4	(link, google)	0.003877

In [144]:
```
fig_dims = (20,8)
fig, ax = plt.subplots(figsize=fig_dims)
sns.set(font_scale=2)
sns.set_style("darkgrid")
palette = sns.set_palette("dark")
ax = sns.barplot(x=bigram1.head(15)['Words'], y=bigram1.head(15)['Freq'],
                 palette=palette)
ax.set(xlabel="Word Pairs",ylabel="Frequency")
plt.ticklabel_format(style='plain',axis='y')
plt.xticks(rotation=70)
plt.title('Top 15 Word Pairs by Frequency')
plt.show()
```



1.6 Keras NN Binary Classification

In [145]:
```
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.utils import to_categorical
from tensorflow.keras import models, layers, optimizers
```

In [146]:
```
model = 0
```

1.6.1 Tokenize Upsampled Tweets

In [147]:
```
tweets = df_upsampled['Tweet']
tokenizer = Tokenizer(num_words=10000)
tokenizer.fit_on_texts(tweets)
sequences = tokenizer.texts_to_sequences(tweets)
print('sequences type: ' , type(sequences))
```

```
sequences type:  <class 'list'>
```

In [148]:
```
one_hot_results = tokenizer.texts_to_matrix(tweets, mode='binary')
print('one_hot_results type:', type(one_hot_results))
one_hot_results = np.asarray(one_hot_results)
```

```
one_hot_results type: <class 'numpy.ndarray'>
```

In [149]:
```
word_index = tokenizer.word_index
print('Found %s unique tokens.' % len(word_index))
```

```
Found 4816 unique tokens.
```

In [150]:
```
print('Dimensions of our coded results:', np.shape(one_hot_results))
```

```
Dimensions of our coded results: (3500, 10000)
```

In [151]:
```
y = df_upsampled['Positive_Bin']
```

In [152]:
```
y = np.asarray(y)
```

In [153]:
```
print(y.shape)
print(one_hot_results.shape)
```

```
(3500,)
(3500, 10000)
```

In [154]:
```
print(len(y))
```

```
3500
```

In [155]:
```
import random
```

In [156]:
```
random.seed(42)
test_index = list(random.sample(range(1,3200),2000))

test = np.asarray(one_hot_results[test_index])
train = np.delete(one_hot_results, test_index, 0)

label_test = y[test_index]
label_train = np.delete(y, test_index, 0)

print('Test label shape:', np.shape(label_test))
print('Train label shape:', np.shape(label_train))
print('Test shape:', np.shape(test))
print('Train shape:', np.shape(train))
```

```
Test label shape: (2000,)
Train label shape: (1500,)
Test shape: (2000, 10000)
Train shape: (1500, 10000)
```

In [157]:
```
tokenizer.word_counts
```

Out[157]:
```
OrderedDict([('at', 1127),
             ('sxsw', 3630),
             ('tapworthy', 44),
             ('ipad', 1213),
             ('design', 89),
             ('headaches', 41),
             ('avoiding', 3),
             ('the', 1847),
             ('pitfalls', 3),
             ('of', 753),
             ('new', 357),
             ('challenges', 3),
             ('rt', 1000),
             ('mention', 2312),
             ('part', 12),
             ('journalsim', 5),
             ('is', 883),
             ('support', 15),
             ('democracy', 5),
```

In [158]:
```
print(type(X),X.shape)
print(type(y),y.shape)
```

```
<class 'pandas.core.series.Series'> (3500,)
<class 'numpy.ndarray'> (3500,)
```

1.6.2 Build Neural Network Model with Sigmoid Activation

In [159]:
```
# Initialize a sequential model
model = []
model = models.Sequential()
# Two layers with relu activation
model.add(layers.Dense(32, activation='relu', input_shape=(10000,)))
model.add(layers.Dense(16, activation='relu'))
model.add(layers.Dense(1, activation='sigmoid'))
model.compile(optimizer='adam',
              loss='binary_crossentropy',
              metrics=['acc'])
```

In [160]:
```
model.summary()
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
dense (Dense)                (None, 32)                320032    
_________________________________________________________________
dense_1 (Dense)              (None, 16)                528       
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 17        
=================================================================
Total params: 320,577
Trainable params: 320,577
Non-trainable params: 0
_________________________________________________________________
```

In [161]:
```
train.shape
```

Out[161]:
```
(1500, 10000)
```

In [162]:
```
label_train.shape
```

Out[162]:
```
(1500,)
```

1.6.3 Run Model

In [163]:

```
history = model.fit(train, label_train, batch_size=32, epochs=20, verbose=2,
                    validation_split=.2)
```

```
Epoch 1/20
38/38 - 2s - loss: 0.6475 - acc: 0.6825 - val_loss: 0.4453 - val_acc: 1.000
0
Epoch 2/20
38/38 - 0s - loss: 0.5181 - acc: 0.7325 - val_loss: 0.3239 - val_acc: 0.976
7
Epoch 3/20
38/38 - 0s - loss: 0.3177 - acc: 0.9108 - val_loss: 0.2295 - val_acc: 0.936
7
Epoch 4/20
38/38 - 0s - loss: 0.1449 - acc: 0.9725 - val_loss: 0.1581 - val_acc: 0.950
0
Epoch 5/20
38/38 - 0s - loss: 0.0655 - acc: 0.9892 - val_loss: 0.1624 - val_acc: 0.930
0
Epoch 6/20
38/38 - 0s - loss: 0.0335 - acc: 0.9967 - val_loss: 0.1459 - val_acc: 0.933
3
Epoch 7/20
38/38 - 0s - loss: 0.0188 - acc: 1.0000 - val_loss: 0.1645 - val_acc: 0.916
7
Epoch 8/20
38/38 - 0s - loss: 0.0114 - acc: 1.0000 - val_loss: 0.1565 - val_acc: 0.920
0
Epoch 9/20
38/38 - 0s - loss: 0.0076 - acc: 1.0000 - val_loss: 0.1623 - val_acc: 0.920
0
Epoch 10/20
38/38 - 0s - loss: 0.0055 - acc: 1.0000 - val_loss: 0.1720 - val_acc: 0.920
0
Epoch 11/20
38/38 - 0s - loss: 0.0042 - acc: 1.0000 - val_loss: 0.1660 - val_acc: 0.920
0
Epoch 12/20
38/38 - 0s - loss: 0.0033 - acc: 1.0000 - val_loss: 0.1752 - val_acc: 0.920
0
Epoch 13/20
38/38 - 0s - loss: 0.0026 - acc: 1.0000 - val_loss: 0.1759 - val_acc: 0.920
0
Epoch 14/20
38/38 - 0s - loss: 0.0021 - acc: 1.0000 - val_loss: 0.1779 - val_acc: 0.920
0
Epoch 15/20
38/38 - 0s - loss: 0.0018 - acc: 1.0000 - val_loss: 0.1817 - val_acc: 0.920
0
Epoch 16/20
38/38 - 0s - loss: 0.0015 - acc: 1.0000 - val_loss: 0.1797 - val_acc: 0.920
0
Epoch 17/20
38/38 - 0s - loss: 0.0013 - acc: 1.0000 - val_loss: 0.1855 - val_acc: 0.920
0
Epoch 18/20
38/38 - 0s - loss: 0.0011 - acc: 1.0000 - val_loss: 0.1845 - val_acc: 0.920
0
```

```
Epoch 19/20
38/38 - 0s - loss: 9.7747e-04 - acc: 1.0000 - val_loss: 0.1834 - val_acc:
0.9200
Epoch 20/20
38/38 - 0s - loss: 8.5869e-04 - acc: 1.0000 - val_loss: 0.1911 - val_acc:
0.9200
```

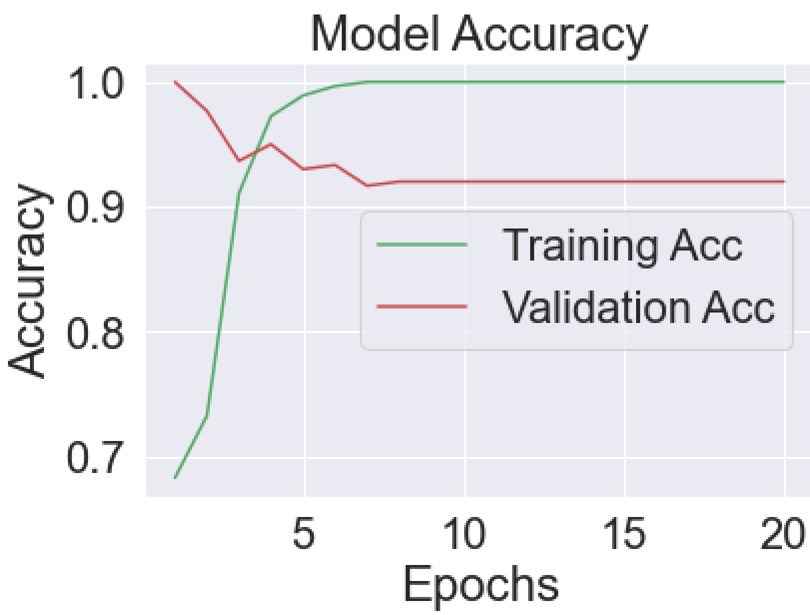
1.6.4 Training and Validation Graphs

In [164]:

```
history_dict = history.history
loss_values = history_dict['loss']
loss_valid = history_dict['val_loss']

epochs = range(1, len(loss_values) + 1)

plt.plot(epochs, loss_values, 'g', label='Training Loss')
plt.plot(epochs, loss_valid, 'r', label='Validation Loss')
plt.title('Training Loss')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.legend()
plt.show()
```

In [165]:

```
# Plot the training accuracy vs the number of epochs

acc_values = history_dict['acc']
acc_valid = history_dict['val_acc']

plt.figure()

plt.plot(epochs, acc_values, 'g', label='Training Acc')
plt.plot(epochs, acc_valid, 'r', label='Validation Acc')
plt.title('Model Accuracy')
plt.xlabel('Epochs')
plt.ylabel('Accuracy')
plt.legend(loc='right')
plt.show()
```



1.7 NLP using Word2Vec

In [166]:

```
from nltk import word_tokenize
```

1.7.1 Tokenize Tweets

In [167]:
```
data = df_upsampled['Tweet'].map(word_tokenize)
```

In [168]:
```
data[:10]
```

Out[168]:
```
1749    [At, #, sxsw, #, tapworthy, iPad, Design, Head...
6436    [RT, @, mention, Part, of, Journalsim, is, the...
3838    [Fuck, the, iphone, !, RT, @, mention, New, #,...
1770    [#, SXSW, 2011, :, Novelty, of, iPad, news, ap...
1062    [New, #, SXSW, rule, :, no, more, ooing, and, ...
324     [Overheard, at, #, sxsw, interactive, :, &, qu...
1944    [#, virtualwallet, #, sxsw, no, NFC, in, #, ip...
7201    [#, SXSW, a, tougher, crowd, than, Colin, Quin...
3159    [Why, is, wifi, working, on, my, laptop, but, ...
4631    [Is, starting, to, think, my, #, blackberry, i...
Name: Tweet, dtype: object
```

1.7.2 Create Word2Vec Model

In [169]:
```
model_W2V = Word2Vec(data, size =100, window=5, min_count=1, workers=4)
```

```
2020-12-30 10:38:26,013 : INFO : collecting all words and their counts
2020-12-30 10:38:26,015 : INFO : PROGRESS: at sentence #0, processed 0 wo
rds, keeping 0 word types
2020-12-30 10:38:26,038 : INFO : collected 5920 word types from a corpus
of 86715 raw words and 3500 sentences
2020-12-30 10:38:26,039 : INFO : Loading a fresh vocabulary
2020-12-30 10:38:26,054 : INFO : effective_min_count=1 retains 5920 uniqu
e words (100% of original 5920, drops 0)
2020-12-30 10:38:26,056 : INFO : effective_min_count=1 leaves 86715 word
corpus (100% of original 86715, drops 0)
2020-12-30 10:38:26,085 : INFO : deleting the raw counts dictionary of 59
20 items
2020-12-30 10:38:26,086 : INFO : sample=0.001 downsamples 52 most-common
words
2020-12-30 10:38:26,087 : INFO : downsampling leaves estimated 56808 word
corpus (65.5% of prior 86715)
2020-12-30 10:38:26,101 : INFO : estimated required memory for 5920 words
and 100 dimensions: 7696000 bytes
2020-12-30 10:38:26,102 : INFO : resetting layer weights
```

In [170]:
```
model_W2V.train(data,total_examples=model_W2V.corpus_count, epochs=10)
```

```
2020-12-30 10:38:27,774 : WARNING : Effective 'alpha' higher than previou
s training cycles
2020-12-30 10:38:27,776 : INFO : training model with 4 workers on 5920 vo
cabulary and 100 features, using sg=0 hs=0 sample=0.001 negative=5 window
=5
2020-12-30 10:38:27,849 : INFO : worker thread finished; awaiting finish
of 3 more threads
2020-12-30 10:38:27,851 : INFO : worker thread finished; awaiting finish
of 2 more threads
2020-12-30 10:38:27,854 : INFO : worker thread finished; awaiting finish
of 1 more threads
2020-12-30 10:38:27,855 : INFO : worker thread finished; awaiting finish
of 0 more threads
2020-12-30 10:38:27,856 : INFO : EPOCH - 1 : training on 86715 raw words
(56838 effective words) took 0.1s, 822211 effective words/s
2020-12-30 10:38:27,916 : INFO : worker thread finished; awaiting finish
of 3 more threads
2020-12-30 10:38:27,925 : INFO : worker thread finished; awaiting finish
of 2 more threads
```

In [171]:
```
wv = model_W2V.wv
```

In [172]:
```
wv.most_similar(positive='phone')
```

```
2020-12-30 10:38:28,533 : INFO : precomputing L2-norms of word weight vecto
rs
```

Out[172]:
```
[('brain', 0.9711115956306458),
 ('makes', 0.9684733152389526),
 ('Double', 0.9683538675308228),
 ('3/20', 0.9668301343917847),
 ('curse', 0.9639092683792114),
 ('words', 0.9631701111793518),
 ('Typing', 0.9614925384521484),
 ('3g', 0.9595977067947388),
 ('Qrank', 0.9587474465370178),
 ('nor', 0.9580039978027344)]
```

In [173]:
```python
wv['help']
```

Out[173]:
```
array([-1.15405507e-01,  1.14294894e-01,  8.07320923e-02, -2.13220775e-01,
       -5.01800716e-01,  2.13985220e-02, -3.78785277e-04,  3.76764052e-02,
       -1.50542289e-01,  4.03295234e-02,  3.25406641e-01,  1.20190106e-01,
        1.00255959e-01,  9.06978622e-02,  8.85991082e-02,  1.57696098e-01,
       -3.51698659e-02,  4.02840137e-01, -2.34701447e-02, -1.46638170e-01,
       -1.14586376e-01, -1.71307534e-01, -1.34530634e-01,  1.70941520e-02,
       -1.62698463e-01, -2.61524379e-01,  4.83386591e-02,  1.38405904e-01,
        3.82361114e-02,  1.36051252e-01, -9.19375569e-02,  2.21098721e-01,
        4.79477316e-01,  1.41353086e-01, -2.88660824e-01,  1.52132332e-01,
       -2.74485320e-01,  1.38044313e-01,  2.72632372e-02,  9.00184289e-02,
       -1.01076765e-02, -6.93503022e-02, -2.30569899e-01,  9.60568637e-02,
        1.47054285e-01, -2.53053635e-01, -1.36887422e-02,  2.93522596e-01,
       -1.19452722e-01,  1.17573209e-01,  3.28345858e-02,  2.13188231e-01,
       -1.58113152e-01, -3.92688960e-01, -1.19782530e-01,  1.47171784e-02,
       -8.62376094e-02,  4.89073157e-01, -5.26246503e-02,  1.51341826e-01,
       -1.32614389e-01, -2.28723101e-02,  3.85659426e-01, -1.17805719e-01,
       -2.11446453e-02,  1.43716991e-01,  1.86417922e-01,  8.90968442e-02,
       -1.47235528e-01, -3.06924582e-01, -3.52315158e-01, -4.88807112e-01,
        2.40349710e-01,  2.76960254e-01, -1.05115332e-01, -4.35439348e-02,
       -1.00465320e-01, -1.32709876e-01,  1.02236226e-01,  2.25574467e-02,
       -7.36548528e-02, -8.65015462e-02,  4.10651207e-01,  4.27667052e-02,
        1.47791591e-03,  4.97815982e-02, -4.72605303e-02, -1.05518542e-01,
        5.78798167e-02, -2.08741706e-02,  2.17183959e-03, -1.79858267e-01,
       -3.78274135e-02,  2.66649604e-01, -1.16654523e-01, -2.13188633e-01,
        1.65979136e-02, -1.39415190e-01,  1.47022754e-01,  2.09375188e-01],
      dtype=float32)
```

In [174]:
```python
wv.vectors
```

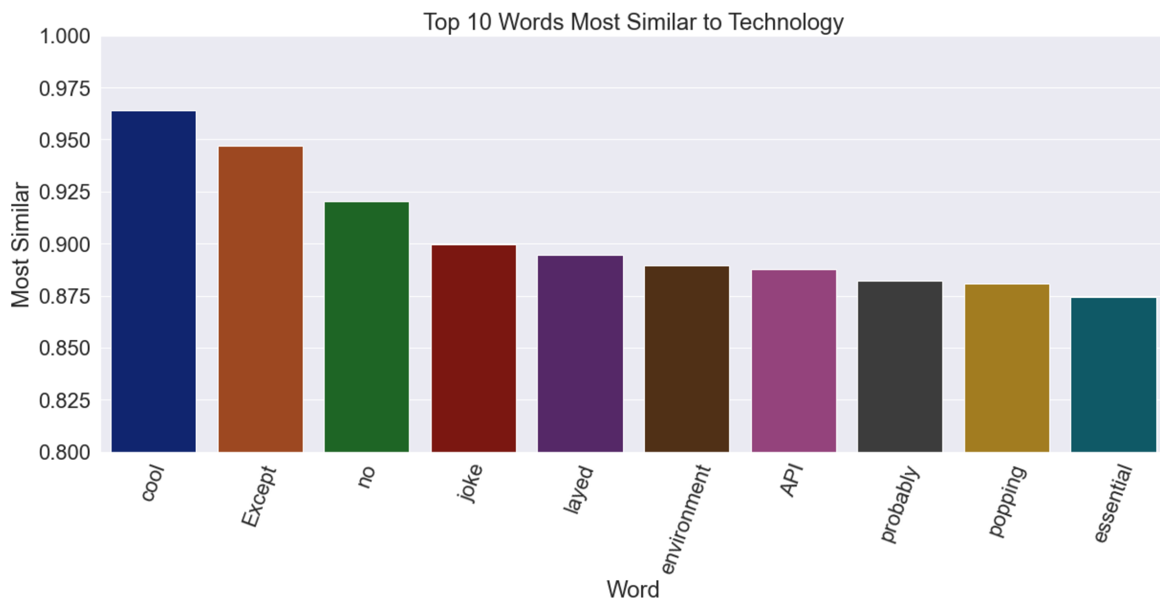
Out[174]:
```
array([[ 2.9242423e-01,  7.4510354e-01,  1.4197147e-01, ...,
        -4.5863187e-01,  2.8571410e-02, -7.1264851e-01],
       [-3.4958524e-01,  7.8811604e-01,  2.9253495e-01, ...,
         1.5906949e-01,  2.3833640e-01,  3.4809700e-01],
       [-5.9380271e-02, -3.9695300e-02, -5.4707265e-01, ...,
         1.3141860e+00,  1.0762907e+00, -1.3292576e+00],
       ...,
       [-1.9461019e-02,  2.0938240e-02,  5.2417060e-03, ...,
        -1.4613664e-02,  6.1093946e-04, -2.2308560e-02],
       [ 2.8513556e-02, -3.8645808e-02, -1.9446509e-02, ...,
        -4.4995070e-02,  1.1583727e-02, -2.3439007e-02],
       [-1.5608729e-02, -4.6526408e-03, -3.2827316e-03, ...,
        -5.6485850e-03,  2.1757590e-02,  4.2654656e-02]], dtype=float32)
```

In [175]:
```python
df_tech = pd.DataFrame(wv.most_similar(positive=['technology']))
```

1.7.3 Most Similar Words

In [176]:

```
fig_dims = (20,8)
fig, ax = plt.subplots(figsize=fig_dims)
sns.set(font_scale=2)
sns.set_style("darkgrid")
palette = sns.set_palette("dark")
ax = sns.barplot(x=df_tech.head(10)[0], y=df_tech.head(10)[1],
                 palette=palette)
ax.set(xlabel="Word",ylabel="Most Similar")
plt.ticklabel_format(style='plain',axis='y')
plt.ylim(.8,1)
plt.xticks(rotation=70)
plt.title('Top 10 Words Most Similar to Technology')
plt.show()
```

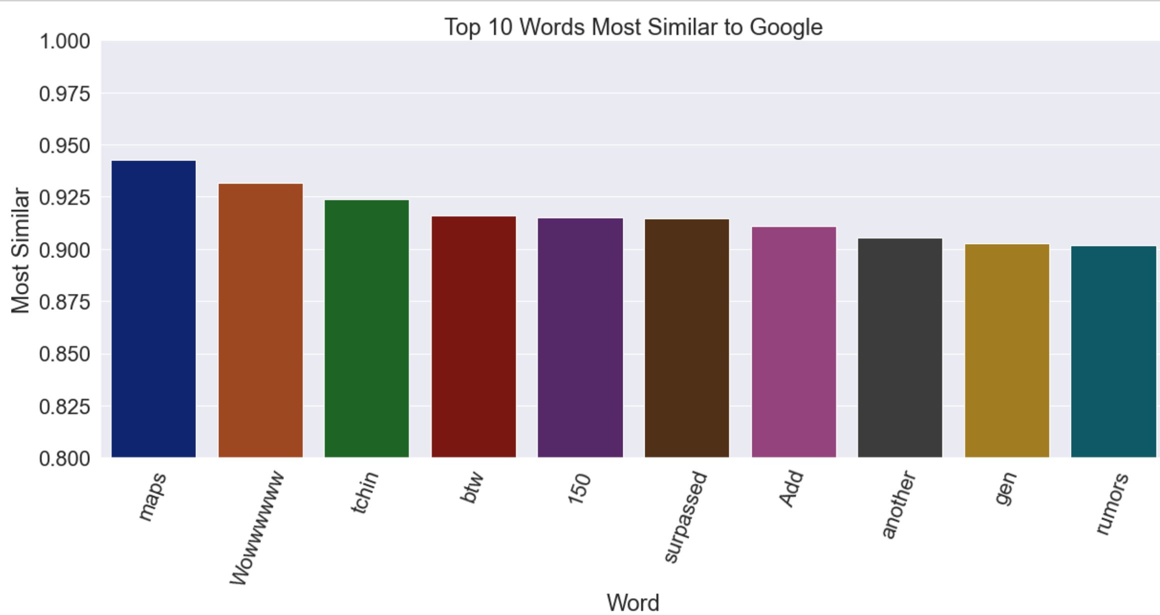


In [177]:
```
df_google = pd.DataFrame(wv.most_similar(positive=['google']))
df_google
```

Out[177]:

	0	1
0	maps	0.942716
1	Wowwwwww	0.931714
2	tchin	0.923837
3	btw	0.916241
4	150	0.915066
5	surpassed	0.914751
6	Add	0.910935
7	another	0.905318
8	gen	0.902506
9	rumors	0.901651

In [178]:
```
fig_dims = (20,8)
fig, ax = plt.subplots(figsize=fig_dims)
sns.set(font_scale=2)
sns.set_style("darkgrid")
palette = sns.set_palette("dark")
ax = sns.barplot(x=df_google.head(10)[0], y=df_google.head(10)[1],
                 palette=palette)
ax.set(xlabel="Word",ylabel="Most Similar")
plt.ticklabel_format(style='plain',axis='y')
plt.ylim(.8,1)
plt.xticks(rotation=70)
plt.title('Top 10 Words Most Similar to Google')
plt.show()
```



In [179]:

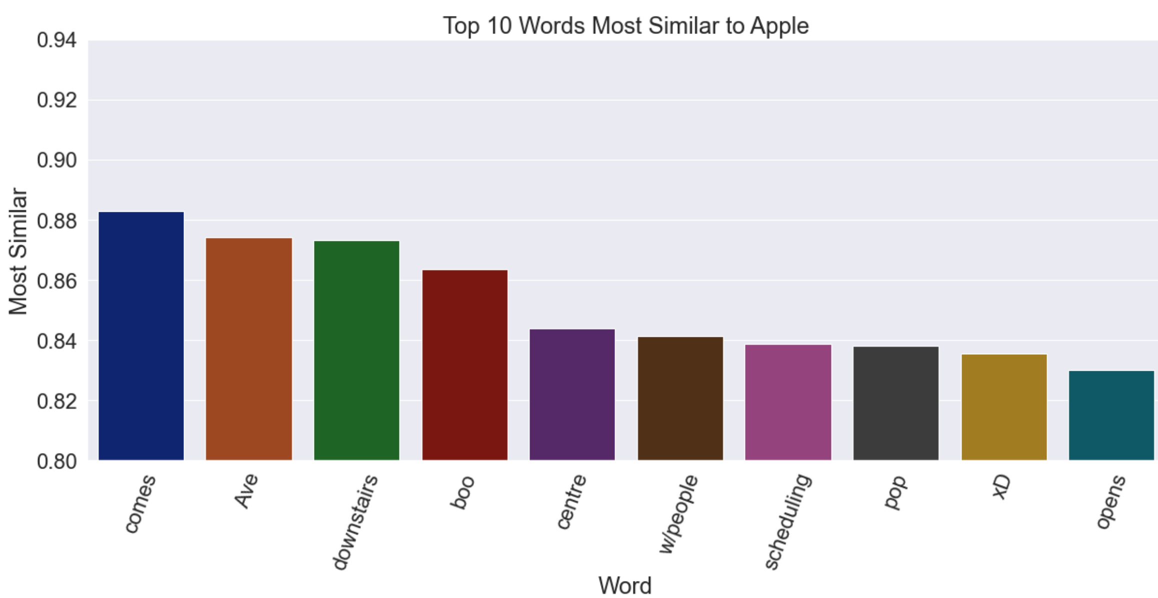
```
df_apple = pd.DataFrame(wv.most_similar(positive=['apple']))
df_apple
```

Out[179]:

	0	1
0	comes	0.883024
1	Ave	0.874297
2	downstairs	0.873119
3	boo	0.863546
4	centre	0.843875
5	w/people	0.841283
6	scheduling	0.838614
7	pop	0.838168
8	xD	0.835714
9	opens	0.830044

In [180]:

```
fig_dims = (20,8)
fig, ax = plt.subplots(figsize=fig_dims)
sns.set(font_scale=2)
sns.set_style("darkgrid")
palette = sns.set_palette("dark")
ax = sns.barplot(x=df_apple.head(10)[0], y=df_apple.head(10)[1], palette=pale
ax.set(xlabel="Word",ylabel="Most Similar")
plt.ticklabel_format(style='plain',axis='y')
plt.ylim(.8,.94)
plt.xticks(rotation=70)
plt.title('Top 10 Words Most Similar to Apple')
plt.show()
```



In [181]:
```
import nltk
nltk.download('vader_lexicon')

import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import random
from sklearn.model_selection import train_test_split
from keras.utils.np_utils import to_categorical
from sklearn import preprocessing
from keras.preprocessing.text import Tokenizer
from keras import models
from keras import layers
from keras import optimizers
```

```
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     C:\Users\josep\AppData\Roaming\nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
```

1.8 Keras NN Multiple Classification

In [182]:
```
df = pd.read_csv('Tweet.csv')
df_up = pd.read_csv('Upsampled.csv')
```

In [183]:
```
df = df.drop(columns='Unnamed: 0')
```

In [184]:
```
df.head(5) # normal
```

Out[184]:

	Tweet	Platform	Emotion	Positive_Bin
0	.@wesley83 I have a 3G iPhone. After 3 hrs twe...	iPhone	Negative emotion	0
1	@jessedee Know about @fludapp ? Awesome iPad/i...	iPad or iPhone App	Positive emotion	1
2	@swonderlin Can not wait for #iPad 2 also. The...	iPad	Positive emotion	1
3	@sxsw I hope this year's festival isn't as cra...	iPad or iPhone App	Negative emotion	0
4	@sxtxstate great stuff on Fri #SXSW: Marissa M...	Google	Positive emotion	1

In [185]:
```
df_up = df_up.drop(columns='Unnamed: 0')
```

In [186]:
```
df_up.head(5) # upsampled for increased number of negative tweets
```

Out[186]:

	Tweet	Platform	Emotion	Positive_Bin
0	At #sxsw #tapworthy iPad Design Headaches - av...	iPad	Negative emotion	0
1	RT @mention Part of Journalsim is the support ...	NaN	Negative emotion	0
2	Fuck the iphone! RT @mention New #UberSocial f...	iPhone	Negative emotion	0
3	#SXSW 2011: Novelty of iPad news apps fades fa...	iPad	Negative emotion	0
4	New #SXSW rule: no more ooing and ahing over y...	iPad	Negative emotion	0

In [187]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3548 entries, 0 to 3547
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   Tweet         3548 non-null   object
 1   Platform      3191 non-null   object
 2   Emotion       3548 non-null   object
 3   Positive_Bin  3548 non-null   int64 
dtypes: int64(1), object(3)
memory usage: 111.0+ KB
```

In [188]:
```
df_up.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3500 entries, 0 to 3499
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   Tweet         3500 non-null   object
 1   Platform      3171 non-null   object
 2   Emotion       3500 non-null   object
 3   Positive_Bin  3500 non-null   int64 
dtypes: int64(1), object(3)
memory usage: 109.5+ KB
```

In [189]:
```
df_up['Positive_Bin'].value_counts()
```

Out[189]:
```
1    2500
0    1000
Name: Positive_Bin, dtype: int64
```

1.8.1 VADER Sentiment Analysis

In [190]:
```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
```

In [191]:
```
sid = SentimentIntensityAnalyzer()
```

In [192]:
```
df_up['scores'] = df_up['Tweet'].apply(lambda review:sid.polarity_scores(revi
```

In [193]:
```
df_up['compound'] = df_up['scores'].apply(lambda d:d['compound'])
```

In [194]:
```
df_up['comp_score'] = df_up['compound'].apply(lambda score: 1
                                              if score >= 0 else 0)
```

In [195]:
```
df_up.head()
```

Out[195]:

	Tweet	Platform	Emotion	Positive_Bin	scores	compound	comp_score
0	At #sxsw #tapworthy iPad Design Headaches - av...	iPad	Negative emotion	0	{'neg': 0.153, 'neu': 0.764, 'pos': 0.083, 'co...	-0.2732	0
1	RT @mention Part of Journalsim is the support ...	NaN	Negative emotion	0	{'neg': 0.0, 'neu': 0.63, 'pos': 0.37, 'compou...	0.8796	1
2	Fuck the iphone! RT @mention New #UberSocial f...	iPhone	Negative emotion	0	{'neg': 0.166, 'neu': 0.834, 'pos': 0.0, 'comp...	-0.5848	0
3	#SXSW 2011: Novelty of iPad news apps fades fa...	iPad	Negative emotion	0	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	0.0000	1
4	New #SXSW rule: no more ooing and ahing over y...	iPad	Negative emotion	0	{'neg': 0.083, 'neu': 0.83, 'pos': 0.087, 'com...	0.0258	1

In [196]:
```
from sklearn.metrics import accuracy_score, classification_report
from sklearn.metrics import confusion_matrix, plot_confusion_matrix
```

In [197]:
```
acc_score = accuracy_score(df_up['Positive_Bin'],df_up['comp_score'])
```

In [198]:
```
print('Accuracy Score: ', "{:.3f}".format(acc_score*100),"%")
```

```
Accuracy Score:  75.371 %
```

In [199]:
```
print(classification_report(df_up['Positive_Bin'],df_up['comp_score']))
```

```
              precision    recall  f1-score   support

           0       0.61      0.39      0.47      1000
           1       0.79      0.90      0.84      2500

    accuracy                           0.75      3500
   macro avg       0.70      0.64      0.66      3500
weighted avg       0.74      0.75      0.73      3500
```

1.8.2 VADER Confusion Matrix

In [200]:
```
confusion_matrix(df_up['Positive_Bin'],df_up['comp_score'])
```

Out[200]:
```
array([[ 389,  611],
       [ 251, 2249]], dtype=int64)
```

VADER doesn't do a great job of correctly classifying tweet sentiment, with 611 false positive tweets that are actually negative

In [201]:
```
full_df = pd.read_csv('Full_DF')
```

In [202]:
```
full_df.head()
```

Out[202]:

	Unnamed: 0	Tweet	Platform	Emotion	Uncertain	Negative	No Emotion	Positive
0	0	.@wesley83 I have a 3G iPhone. After 3 hrs twe...	iPhone	Negative emotion	0	1	0	0
1	1	@jessedee Know about @fludapp ? Awesome iPad/i...	iPad or iPhone App	Positive emotion	0	0	0	1
2	2	@swonderlin Can not wait for #iPad 2 also. The...	iPad	Positive emotion	0	0	0	1
3	3	@sxsw I hope this year's festival isn't as cra...	iPad or iPhone App	Negative emotion	0	1	0	0
4	4	@sxtxstate great stuff on Fri #SXSW: Marissa M...	Google	Positive emotion	0	0	0	1

In [203]:
```
full_df = full_df.drop(columns='Unnamed: 0')
```

In [204]:
```
full_df.head(10)
full_df = full_df.dropna()
```

1.8.3 Tokenize Tweets

In [205]:

```
tweets = full_df['Tweet']
tokenizer = Tokenizer(num_words=5000)
tokenizer.fit_on_texts(tweets)
sequences = tokenizer.texts_to_sequences(tweets)
print('sequences type: ' , type(sequences))
```

```
sequences type:  <class 'list'>
```

In [206]:

```
one_hot_results = tokenizer.texts_to_matrix(tweets, mode='binary')
print('one_hot_results type:', type(one_hot_results))
```

```
one_hot_results type: <class 'numpy.ndarray'>
```

In [207]:

```
word_index = tokenizer.word_index
print('Found %s unique tokens.' % len(word_index))
```

```
Found 5963 unique tokens.
```

In [208]:

```
# Our coded data
print('Dimensions of our coded results:', np.shape(one_hot_results))
```

```
Dimensions of our coded results: (3291, 5000)
```

In [209]:

```
print(y.shape)
print(one_hot_results.shape)
```

```
(3500,)
(3291, 5000)
```

In [210]:

```
emotion = full_df['Emotion']

# Initialize
le = preprocessing.LabelEncoder()
le.fit(emotion)
print('Original class labels:')
print(list(le.classes_))
print('\n')
emotion_cat = le.transform(emotion)

# If you wish to retrieve the original descriptive labels post production
# list(le.inverse_transform([0, 1, 3, 3, 0, 6, 4]))

print('New product labels:')
print(emotion_cat)
print('\n')

# Each row will be all zeros except for the category for that observation
print('One hot labels; 4 binary columns, one for each of the categories.')
product_onehot = to_categorical(emotion_cat)
print(product_onehot)
print('\n')

print('One hot labels shape:')
print(np.shape(product_onehot))
```

```
Original class labels:
["I can't tell", 'Negative emotion', 'No emotion toward brand or product',
'Positive emotion']


New product labels:
[1 3 3 ... 1 3 3]


One hot labels; 4 binary columns, one for each of the categories.
[[0. 1. 0. 0.]
 [0. 0. 0. 1.]
 [0. 0. 0. 1.]
 ...
 [0. 1. 0. 0.]
 [0. 0. 0. 1.]
 [0. 0. 0. 1.]]


One hot labels shape:
(3291, 4)
```

In [211]:
```
random.seed(42)
test_index = random.sample(range(1,3200), 1500)

test = one_hot_results[test_index]
train = np.delete(one_hot_results, test_index, 0)

label_test = product_onehot[test_index]
label_train = np.delete(product_onehot, test_index, 0)

print('Test label shape:', np.shape(label_test))
print('Train label shape:', np.shape(label_train))
print('Test shape:', np.shape(test))
print('Train shape:', np.shape(train))
```

```
Test label shape: (1500, 4)
Train label shape: (1791, 4)
Test shape: (1500, 5000)
Train shape: (1791, 5000)
```

1.8.4 Build Neural Network Model

In [212]:
```
from keras.layers import Input, Dense, LSTM, Embedding
from keras.layers import Dropout, Activation, Bidirectional, GlobalMaxPool1D
from keras.models import Sequential
```

In [213]:
```
# Initialize  and build a sequential model
model = models.Sequential()
# Two layers with relu activation
model.add(layers.Dense(50, activation='relu', input_shape=(5000,)))
model.add(layers.Dense(25, activation='relu'))
model.add(layers.Dense(4, activation='softmax'))
model.compile(optimizer='adam',
              loss='categorical_crossentropy',
              metrics=['acc'])
```

1.8.5 Run Model

In [214]:
```
history = model.fit(train,
                    label_train,
                    epochs=20,
                    batch_size=32,
                    validation_split=.2)
```

```
Epoch 1/20
45/45 [==============================] - 1s 11ms/step - loss: 1.1329 - ac
c: 0.7339 - val_loss: 0.6482 - val_acc: 0.8162
Epoch 2/20
45/45 [==============================] - 0s 4ms/step - loss: 0.5271 - ac
c: 0.8037 - val_loss: 0.5969 - val_acc: 0.8134
Epoch 3/20
45/45 [==============================] - 0s 3ms/step - loss: 0.3266 - ac
c: 0.8925 - val_loss: 0.5757 - val_acc: 0.8357
Epoch 4/20
45/45 [==============================] - 0s 4ms/step - loss: 0.1697 - ac
c: 0.9527 - val_loss: 0.5984 - val_acc: 0.8384
Epoch 5/20
45/45 [==============================] - 0s 4ms/step - loss: 0.1019 - ac
c: 0.9660 - val_loss: 0.6322 - val_acc: 0.8412
Epoch 6/20
45/45 [==============================] - 0s 4ms/step - loss: 0.0746 - ac
c: 0.9744 - val_loss: 0.6516 - val_acc: 0.8273
Epoch 7/20
45/45 [==============================] - 0s 4ms/step - loss: 0.0518 - ac
c: 0.9889 - val_loss: 0.7120 - val_acc: 0.8329
Epoch 8/20
45/45 [==============================] - 0s 4ms/step - loss: 0.0269 - ac
c: 0.9938 - val_loss: 0.7488 - val_acc: 0.8384
Epoch 9/20
45/45 [==============================] - 0s 4ms/step - loss: 0.0250 - ac
c: 0.9937 - val_loss: 0.7861 - val_acc: 0.8357
Epoch 10/20
45/45 [==============================] - 0s 4ms/step - loss: 0.0230 - ac
c: 0.9947 - val_loss: 0.8327 - val_acc: 0.8412
Epoch 11/20
45/45 [==============================] - 0s 4ms/step - loss: 0.0143 - ac
c: 0.9970 - val_loss: 0.8769 - val_acc: 0.8412
Epoch 12/20
45/45 [==============================] - 0s 5ms/step - loss: 0.0169 - ac
c: 0.9937 - val_loss: 0.8605 - val_acc: 0.8329
Epoch 13/20
45/45 [==============================] - 0s 4ms/step - loss: 0.0121 - ac
c: 0.9955 - val_loss: 0.9023 - val_acc: 0.8384
Epoch 14/20
45/45 [==============================] - 0s 4ms/step - loss: 0.0147 - ac
c: 0.9937 - val_loss: 0.9005 - val_acc: 0.8245
Epoch 15/20
45/45 [==============================] - 0s 4ms/step - loss: 0.0139 - ac
c: 0.9944 - val_loss: 0.9751 - val_acc: 0.8412
Epoch 16/20
45/45 [==============================] - 0s 4ms/step - loss: 0.0122 - ac
c: 0.9966 - val_loss: 0.9334 - val_acc: 0.8134
Epoch 17/20
45/45 [==============================] - 0s 4ms/step - loss: 0.0090 - ac
c: 0.9957 - val_loss: 0.9934 - val_acc: 0.8412
```

```
Epoch 18/20
45/45 [==============================] - 0s 4ms/step - loss: 0.0092 - ac
c: 0.9981 - val_loss: 0.9536 - val_acc: 0.8217
Epoch 19/20
45/45 [==============================] - 0s 3ms/step - loss: 0.0080 - ac
c: 0.9976 - val_loss: 0.9951 - val_acc: 0.8273
Epoch 20/20
45/45 [==============================] - 0s 3ms/step - loss: 0.0127 - ac
c: 0.9961 - val_loss: 1.0090 - val_acc: 0.8329
```

In [215]:
```
history_dict = history.history
```

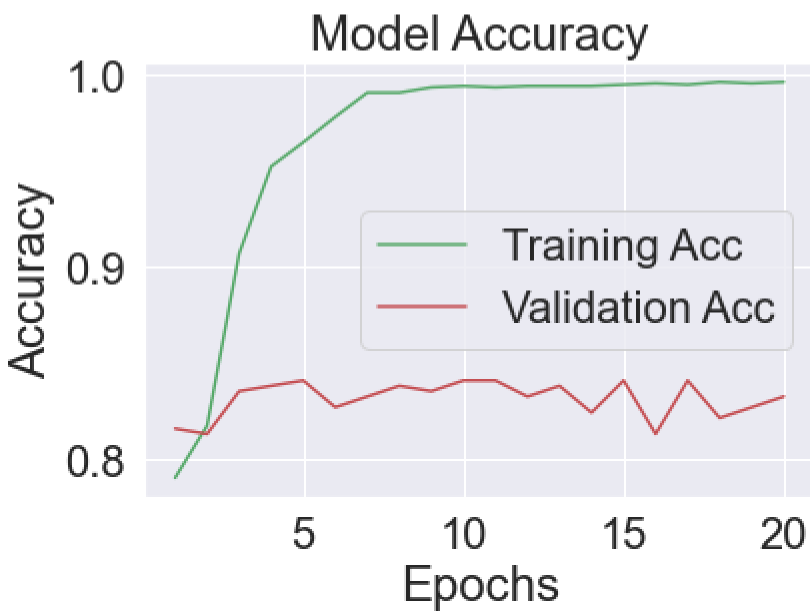
In [216]:
```
history_dict.keys()
```

Out[216]:
```
dict_keys(['loss', 'acc', 'val_loss', 'val_acc'])
```

1.8.6 Training and Validation Graphs

In [217]:
```
history_dict = history.history
loss_values = history_dict['loss']
loss_valid = history_dict['val_loss']

epochs = range(1, len(loss_values) + 1)

plt.plot(epochs, loss_values, 'g', label='Training Loss')
plt.plot(epochs, loss_valid, 'r', label='Validation Loss')
plt.title('Training Loss')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.legend()
plt.show()
```

In [218]:
```
# Plot the training accuracy vs the number of epochs

acc_values = history_dict['acc']
acc_valid = history_dict['val_acc']

plt.figure()

plt.plot(epochs, acc_values, 'g', label='Training Acc')
plt.plot(epochs, acc_valid, 'r', label='Validation Acc')
plt.title('Model Accuracy')
plt.xlabel('Epochs')
plt.ylabel('Accuracy')
plt.legend(loc='right')
plt.show()
```



In [219]:
```
# Output (probability) predictions for the test set
y_hat_test = model.predict(test)
```

In [220]:
```
# Print the loss and accuracy for the training set
results_train = model.evaluate(train, label_train)
results_train
```

```
56/56 [==============================] - 0s 1ms/step - loss: 0.2077 - acc:
0.9648
```

Out[220]:
```
[0.20773755013942719, 0.9648241400718689]
```

In [221]:
```
results_test = model.evaluate(test, label_test)
results_test # model predicts on the test data with almost 84% accuracy.
```

```
47/47 [==============================] - 0s 1ms/step - loss: 0.8366 - acc:
0.8427
```

Out[221]:
```
[0.8366084694862366, 0.8426666855812073]
```

1.9 Question 1 and Recommendation

1.9.1 In tweets targeting either the iPhone or Android phones, which product is more often the subject of negatively charged emotions?

In [222]:
```
df_neg = pd.read_csv('Full_DF')
df_neg = df_neg.drop(columns='Unnamed: 0')
```

In [223]:
```
df_neg.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9093 entries, 0 to 9092
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Tweet       9092 non-null   object
 1   Platform    3291 non-null   object
 2   Emotion     9093 non-null   object
 3   Uncertain   9093 non-null   int64 
 4   Negative    9093 non-null   int64 
 5   No Emotion  9093 non-null   int64 
 6   Positive    9093 non-null   int64 
dtypes: int64(4), object(3)
memory usage: 497.4+ KB
```

In [224]:
```
df_grouped = df_neg.groupby(by=df_neg['Platform']).sum()
```

In [225]:
```
df_grouped.index
```

Out[225]:
```
Index(['Android', 'Android App', 'Apple', 'Google',
       'Other Apple product or service', 'Other Google product or service',
       'iPad', 'iPad or iPhone App', 'iPhone'],
      dtype='object', name='Platform')
```

In [226]:
```
df_grouped
```

Out[226]:

Platform	Uncertain	Negative	No Emotion	Positive
Android	0	8	1	69
Android App	0	8	1	72
Apple	2	95	21	543
Google	1	68	15	346
Other Apple product or service	0	2	1	32
Other Google product or service	1	47	9	236
iPad	4	125	24	793
iPad or iPhone App	0	63	10	397
iPhone	1	103	9	184

In [227]:
```
# separate tweets
df_android = df_grouped.loc[df_grouped.index =='Android']
df_iphone = df_grouped.loc[df_grouped.index =='iPhone']
```

In [228]:
```
df_android
```

Out[228]:

Platform	Uncertain	Negative	No Emotion	Positive
Android	0	8	1	69

In [229]:
```
percent_negative_android_tweets = df_android['Negative']/sum(df_android['Posi
```

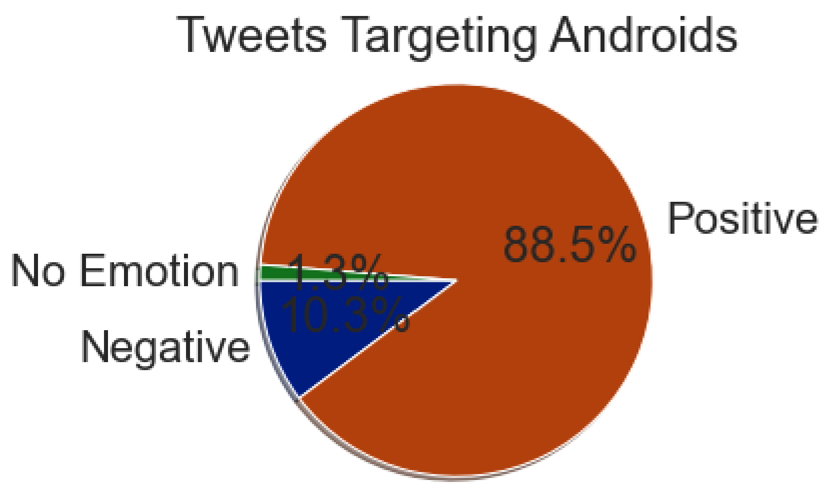
In [230]:
```
print("Percentage of tweets targeting Android phones that are negative: {:.3f
```

```
Percentage of tweets targeting Android phones that are negative: 10.390 %
```

In [231]:
```
labels1 = 'Negative', 'Positive', 'No Emotion'
sizes1 = [8, 69, 1]
```

1.9.2 Negative Tweets

In [232]:
```
fig1, ax1 = plt.subplots()
ax1.pie(sizes1, labels=labels1, autopct='%1.1f%%',
        shadow=True, startangle=180)
ax1.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle
plt.title("Tweets Targeting Androids")

plt.show()
```



In [233]:
```
df_iphone
```

Out[233]:

	Uncertain	Negative	No Emotion	Positive
Platform				
iPhone	1	103	9	184

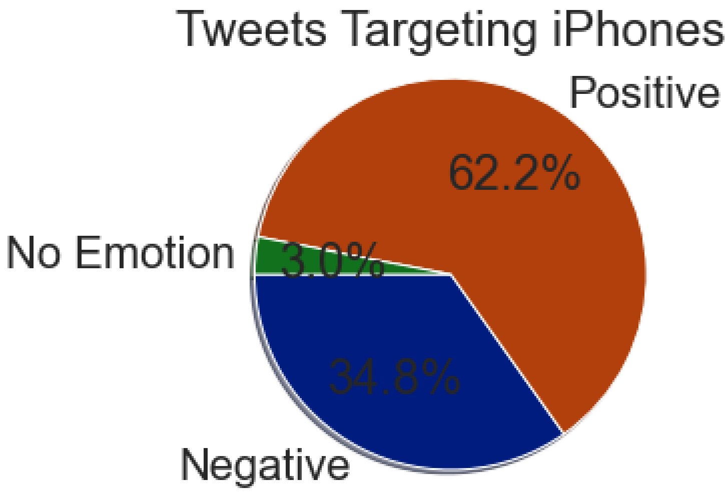
In [234]:
```
percent_negative_iphone_tweets = df_iphone['Negative']/sum(df_iphone['Negativ
```

In [235]:
```
print("Percentage of tweets targeting iPhones that are negative: {:.3f}".form
```

```
Percentage of tweets targeting iPhones that are negative: 35.889 %
```

In [267]:
```
sizes2 = [103, 184, 9]
labels2 = 'Negative', 'Positive', 'No Emotion'
```

In [268]:
```
fig1, ax1 = plt.subplots()
ax1.pie(sizes2, labels=labels2, autopct='%1.1f%%',
        shadow=True, startangle=180)
ax1.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle
plt.title("Tweets Targeting iPhones")

plt.show()
```



1.9.3 Recommendation

In creating your phone, more users may want the option to have a more customizable user interface which the Android provides. We will need to look into more detail about what negative words users are including in their negative tweets that target iPhones to specifically determine users' complaints.

1.10 Question 2 and Recommendation

1.10.1 What words are most common in negative tweets about iPhones and Android phones?

1.10.2 Negative Android Sentiment

In [238]:
```
# collect all negative tweets for each product
df_neg_android = df_neg.loc[df_neg['Platform'] =='Android']
df_neg_iphone = df_neg.loc[df_neg['Platform'] =='iPhone']
```

In [239]:
```
df_neg_android = df_neg_android.loc[df_neg_android['Negative'] == 1]
```

In [240]:
```
df_neg_android # tweets about Android that are negative - create bag of words
```

Out[240]:

	Tweet	Platform	Emotion	Uncertain	Negative	No Emotion	Positive
350	they took away the lego pit but replaced it wi...	Android	Negative emotion	0	1	0	0
1940	Why does all the #Android meetups here in #Aus...	Android	Negative emotion	0	1	0	0
1999	@mention Android needs a way to group apps lik...	Android	Negative emotion	0	1	0	0
3389	Lunch with @mention at #CNNGrill. View from th...	Android	Negative emotion	0	1	0	0
4865	Excited to meet the @mention at #sxsw so I can...	Android	Negative emotion	0	1	0	0
8053	Spending some time this morning resetting my a...	Android	Negative emotion	0	1	0	0
8258	Is it just me or has the @mention client for A...	Android	Negative emotion	0	1	0	0
8801	Auntie's voxpop of popular #sxsw apps is wort...	Android	Negative emotion	0	1	0	0

In [241]:
```
corpus_android = list(df_neg_android['Tweet'])
```

In [242]:
```
corpus_android[:10] # entirety of negative android tweets
```

Out[242]:
```
['they took away the lego pit but replaced it with a recharging station ;)
#sxsw and i might check prices for an iphone - crap samsung android',
 "Why does all the #Android meetups here in #Austin are when I'm at work. W
ell at least there is the PS meetup #sxsw",
 '@mention Android needs a way to group apps like you can now do with iPad/
iPod. #SXSW #hhrs',
 'Lunch with @mention at #CNNGrill. View from the HTML5 dev trenches: Andro
id is painful, iOS is sleek (for what @mention is doing) #sxsw',
 'Excited to meet the @mention at #sxsw so I can show them my Sprint Galaxy
S still running Android 2.1.   #fail',
 'Spending some time this morning resetting my android phone.  First day of
#sxsw was too much for it.',
 'Is it just me or has the @mention client for Android gotten really buggy
lately? #SXSW to blame?',
 "Auntie's voxpop of  popular #sxsw apps is worth a watch: {link} Not many
Android phones on view."]
```

In [243]:
```
# tokenize
android_tokens = word_tokenize(','.join(str(v) for v in corpus_android))

# remove stopwords
stopped_android_tokens = [word.lower() for word in android_tokens if word.low
                          not in stopword_list]
```

In [244]:
```
freq = FreqDist(stopped_android_tokens)
```

In [245]:
```
freq.most_common(25)
```

Out[245]:
```
[('android', 8),
 ('apps', 2),
 ('view', 2),
 ('took', 1),
 ('away', 1),
 ('lego', 1),
 ('pit', 1),
 ('replaced', 1),
 ('recharging', 1),
 ('station', 1),
 ('check', 1),
 ('prices', 1),
 ('iphone', 1),
 ('crap', 1),
 ('samsung', 1),
 ('meetups', 1),
 ('austin', 1),
 ('work', 1),
 ('ps', 1),
 ('meetup', 1),
 ('needs', 1),
 ('way', 1),
 ('group', 1),
 ('like', 1),
 ('ipad/ipod', 1)]
```

1.10.3 Negative iPhone Sentiment

In [246]:
```
df_neg_iphone = df_neg_iphone.loc[df_neg_iphone['Negative'] == 1]
```

In [247]:
```
df_neg_iphone # tweets about iphone that are negative - create bag of words i
```

Out[247]:

	Tweet	Platform	Emotion	Uncertain	Negative	No Emotion	Positive
0	.@wesley83 I have a 3G iPhone. After 3 hrs twe...	iPhone	Negative emotion	0	1	0	0
17	I just noticed DST is coming this weekend. How...	iPhone	Negative emotion	0	1	0	0
92	What !?!? @mention #SXSW does not provide iPh...	iPhone	Negative emotion	0	1	0	0
233	If iPhone alarms botch the timechange, how man...	iPhone	Negative emotion	0	1	0	0
236	I meant I also wish I at #SXSW #dyac stupid i...	iPhone	Negative emotion	0	1	0	0
...	...	...	...	...	...	...	...

In [248]:
```
corpus_iphone = list(df_neg_iphone['Tweet'])
```

In [249]:
```
corpus_iphone[:15]
```

Out[249]:
```
['.@wesley83 I have a 3G iPhone. After 3 hrs tweeting at #RISE_Austin, it w
as dead!  I need to upgrade. Plugin stations at #SXSW.',
 'I just noticed DST is coming this weekend. How many iPhone users will be
an hour late at SXSW come Sunday morning? #SXSW #iPhone',
 "What !?!? @mention  #SXSW does not provide iPhone chargers?!?  I've chang
ed my mind about going next year!",
 "If iPhone alarms botch the timechange, how many #SXSW'ers freak? Late to
flights, missed panels, behind on bloody marys...",
 'I meant I also wish I  at #SXSW #dyac stupid iPhone!',
 'Overheard at #sxsw interactive: &quot;Arg! I hate the iphone! I want my b
lackberry back&quot; #shocked',
 "overheard at MDW (and I'll second it) &quot;halfway through my iPhone bat
tery already and I haven't even boarded the plane to #sxsw&quot; #amateurho
ur",
 "My iPhone battery can't keep up with my tweets! Thanks Apple.  #SXSW  #pr
ecommerce",
 'IPhone is dead. Find me on the secret batphone #sxsw.',
 'Austin is getting full, and #SXSW is underway.  I can tell because my iPh
one is an intermittent brick. #crowded',
 '.@mention I have a 3G iPhone. After 3 hrs tweeting at #RISE_Austin, it wa
s dead!  I need to upgrade. Plugin stations at #SXSW.',
 'my iPhone is overheating. why are there so many british sounding people i
n texas? #SXSW',
 'My iPhone is wilting under the stress of being at #sxsw.',
 'iPhone, I know this #SXSW week will be tough on your already-dwindling ba
ttery, but so help me Jeebus if you keep correcting my curse words.',
 "God, it's like being at #sxsw - have iMac, MacBook, iPhone and BlackBerry
all staring at me. Enough! Time to read a book - remember those?"]
```

In [250]:
```
# tokenize
iphone_tokens = word_tokenize(','.join(str(v) for v in corpus_iphone))

# remove stopwords
stopped_iphone_tokens = [word.lower() for word in iphone_tokens if word.lower
                         not in stopword_list]
```

In [251]:
```
freq = FreqDist(stopped_iphone_tokens)
```

In [252]:
```
freq.most_common(25)
```

Out[252]:
```
[('iphone', 104),
 ('quot', 22),
 ('battery', 15),
 ('amp', 10),
 ('blackberry', 8),
 ('link', 8),
 ('austin', 7),
 ('app', 7),
 ('users', 6),
 ('going', 6),
 ('time', 6),
 ('sxsw.', 5),
 ('like', 5),
 ('u', 5),
 ('good', 5),
 ('3g', 4),
 ('hour', 4),
 ('apple', 4),
 ('people', 4),
 ('know', 4),
 ('ipad', 4),
 ('t-mobile', 4),
 ('shit', 4),
 ('long', 4),
 ('technology', 4)]
```

1.10.4 Recommendation

The Android operating system was claimed to be buggy in addition to someone saying Android is painful and not sleek like Apple's iOS. Generally, users had less negative things to say as a percentage of total comments.

The iPhone was said to have failing battery or a battery charge that does not last long enough when the phone is in operation. Additionally, lack of signal became a problem in crowded areas but this is not typically a phone design issue but instead is an infrastructure problem.

Build a sleek phone with a simple to use Graphical User Interface. Have plenty of battery to power the phone for longer periods. Users would enjoy a feature like a backup battery, or a sleekly designed case that provides a full second charge without adding much volume.

1.11 Question 3 and Recommendation

1.11.1 What are some of the positive features commented about for both iPhones and Android phones?

In [253]:
```
df_pos = pd.read_csv('Full_DF')
```

In [254]:
```
df_pos_android = df_pos.loc[df_pos['Platform'] =='Android']
df_pos_iphone = df_pos.loc[df_pos['Platform'] =='iPhone']
```

In [255]:
```
df_pos_android = df_pos_android.loc[df_pos_android['Positive']==1]
df_pos_iphone = df_pos_iphone.loc[df_pos_iphone['Positive']==1]
```

1.11.2 Positive Android Sentiment

In [256]:
```
corpus_android = list(df_pos_android['Tweet'])
```

In [257]:
```
corpus_android[:20]
```

Out[257]:
```
['#SXSW is just starting, #CTIA is around the corner and #googleio is only
a hop skip and a jump from there, good time to be an #android fan',
 'Excited to meet the @samsungmobileus at #sxsw so I can show them my Sprin
t Galaxy S still running Android 2.1.   #fail',
 'This is a #WINNING picture #android #google #sxsw {link}',
 "I knew if I plied @mention with beer and stogies last night I'd weasel my
way into the Team Android party tonight. #success #SXSW.",
 'Alert the media.  I just saw the one and only Android tablet at #sxsw.  L
ike finding a needle in a haystack!  I also saw a Cr-48.',
 'Farooqui: Now about mobile. iOS, with Android catching up fast and will g
row more once they allow in-app purchasing. #gamesfortv #sxsw',
 'I need to play this game on my #android - #SXSW {link}',
 'Talked to some great developers at the Android meetup. Looking forward to
working with them. #sxsw #android #androidsxsw',
 "There are thousands of iPad 2's floating around Austin at #sxsw and I hav
e not seen even one single Android tablet.  Not even one. Zero.",
 'Woot! RT @mention First Android @mention disc {link} ... Market version c
oming soon! #SXSW',
 'Heard at #sxsw #Android is now the leading market share of smart phones i
n US. #getjarsxsw',
 'Quadroid = Qualcomm + Android just called the platform of the next decade
vs Wintel #sxsw #cloud',
 '{link} via @mention pretty neat database I must say.  does it work on my
#android we shall see. #sxsw #party #free',
 "@mention Android just got a big call out at #sxsw in they #gamelayer open
ing keynote. I knew you'd appreciate.",
 'Android party #sxsw (@mention Lustre Pearl Bar w/ 36 others) {link}',
 '@mention at Team Android party. @mention @mention just walked in. DL Appo
licious app &amp; enter to win free Nexus S! #androidsxsw #sxsw',
 'Piece of awesomeness: Arduino + android = Flaming skulls  {link} @mention
@mention #sxsw #smartthings',
 '@mention Congratulations on winning the Android award! :) #sxsw',
 '@mention crew ripped up Android party - thanks for having us Droid! {lin
k} #sxsw',
 'Great UI demo of @mention on @mention {link} #xoom #sxsw #android #tech #
tablet']
```

In [273]:
```
# tokenize
android_tokens = word_tokenize(','.join(str(v) for v in corpus_android))

# remove stopwords
stopped_android_tokens = [word.lower() for word in android_tokens if word.low
                          not in stopword_list]
```

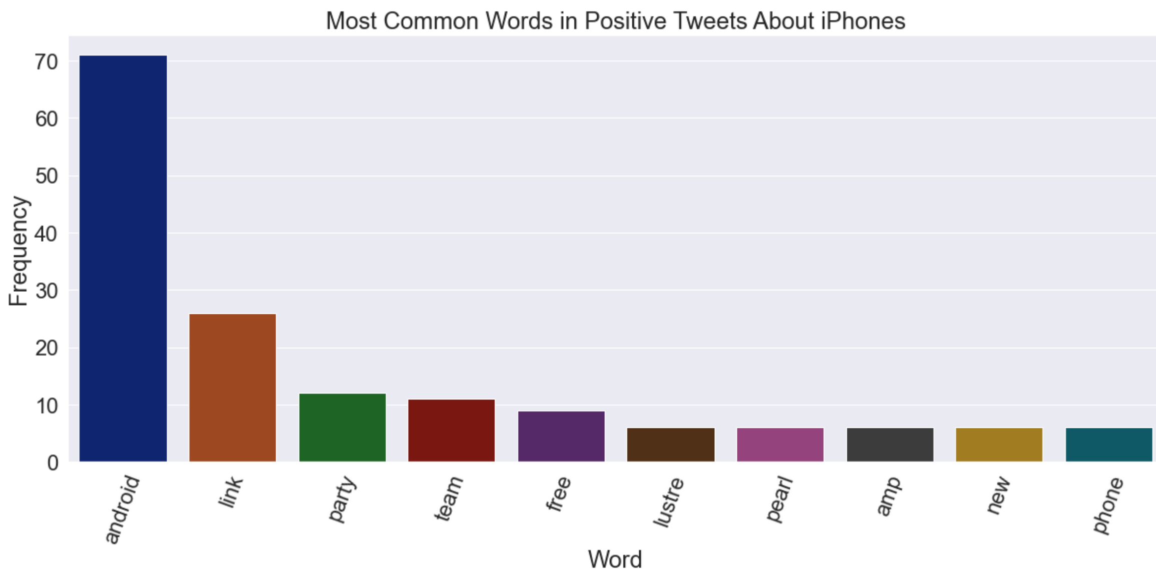
In [274]:
```
freq_and = FreqDist(stopped_android_tokens)
```

In [275]:
```
freq_and.most_common(25)
```

Out[275]:
```
[('android', 71),
 ('link', 26),
 ('party', 12),
 ('team', 11),
 ('free', 9),
 ('lustre', 6),
 ('pearl', 6),
 ('amp', 6),
 ('new', 6),
 ('phone', 6),
 ('dev', 5),
 ('tablet', 4),
 ('need', 4),
 ('great', 4),
 ('meetup', 4),
 ('androidsxsw', 4),
 ('market', 4),
 ('win', 4),
 ('love', 4),
 ('details', 4),
 ('starting', 3),
 ('good', 3),
 ('fan', 3),
 ('excited', 3),
 ('beer', 3)]
```

In [278]:
```
freq_and = pd.DataFrame(freq_and.most_common(25))
```

In [279]:
```
fig_dims = (20,8)
fig, ax = plt.subplots(figsize=fig_dims)
sns.set(font_scale=2)
sns.set_style("darkgrid")
palette = sns.set_palette("dark")
ax = sns.barplot(x=freq_and.head(10)[0], y=freq_and.head(10)[1], palette=pale
ax.set(xlabel="Word",ylabel="Frequency")
plt.ticklabel_format(style='plain',axis='y')
plt.xticks(rotation=70)
plt.title('Most Common Words in Positive Tweets About Android Phones')
plt.show()
```



1.11.3 Positive iPhone Sentiment

In [261]:
```
corpus_iphone = list(df_pos_iphone['Tweet'])
```

In [262]:
```
corpus_iphone[:20]
```

Out[262]:
```
["I love my @mention iPhone case from #Sxsw but I can't get my phone out of
it #fail",
 'Yai!!! RT @mention New #UberSocial for #iPhone now in the App Store inclu
des UberGuide to #SXSW sponsored by (cont) {link}',
 'Take that #SXSW ! RT @mention Major South Korean director gets $130,000 t
o make a movie entirely with his iPhone. {link}',
 'Behind on 100s of emails? Give them all 1 line iPhone composed replies. #
SXSW #protip',
 'Picked up a Mophie battery case 4 my iPhone in prep for #SXSW. Not luggin
g around a laptop &amp; only using my phone was a huge win last year.',
 "Do I need any more for #sxsw! ipad, iphone, laptop, dictaphone, vid.camer
a.... Wow! Love to  meet the REAL 'cerebellum' charged people:)",
 'My iPhone battery at 100%. #winning at #SXSW',
 'BEST SWAG EVER. Thanks @mention My charging iPhone thanks you, too. #SXSW
{link}',
 'Love that I have a MacBook, iPad, and iPhone with me at #sxsw this year.
One runs out of juice, and I can jump to the next.',
 'Holy cow! I just got hooked by Paolo and Alex with a backup charger for m
y iPhone! facebook.com/powermat #powermatteam #sxsw #thanks',
 'Holy cow! I just got hooked by Paolo and Alex with a backup charger for m
y iPhone! facebook.com/powermat #powermattteam #sxsw #thanks',
 "@mention  I'm beyond frustrated w/ @mention after this Samsung Moment run
around &amp; am leaving for ATT &amp; iPhone so I can enjoy #sxsw.",
 "Tim Soo's invisible instruments are jaw dropping. iPhone+Wii controller.
{link} #lovemusicapi #sxsw",
 'I fear no iphone + #att 3gs slowpoke network during #sxsw &amp; #sxswmusi
c.',
 'Check out iPhone Developer Meet Up at SXSW.\n{link} #SXSW',
 "&quot;the iPhone is a transient device used in short bursts; the iPad is
an 'after 8pm, on the couch' device.&quot; @mention #sxsw",
 '@mention  iPhone. Clearly. Positively. Happily. #SXSW',
 'Flipboard is developing an iPhone version, not Android, says @mention #sx
sw',
 "So {link} is part of my presentation at #SXSW so good thing it's crashing
now instead of then. Works best on iPhone/Android",
 'Loving my Morphie JuicePack today for a recharge of iPhone. So worth it.
#sxsw']
```

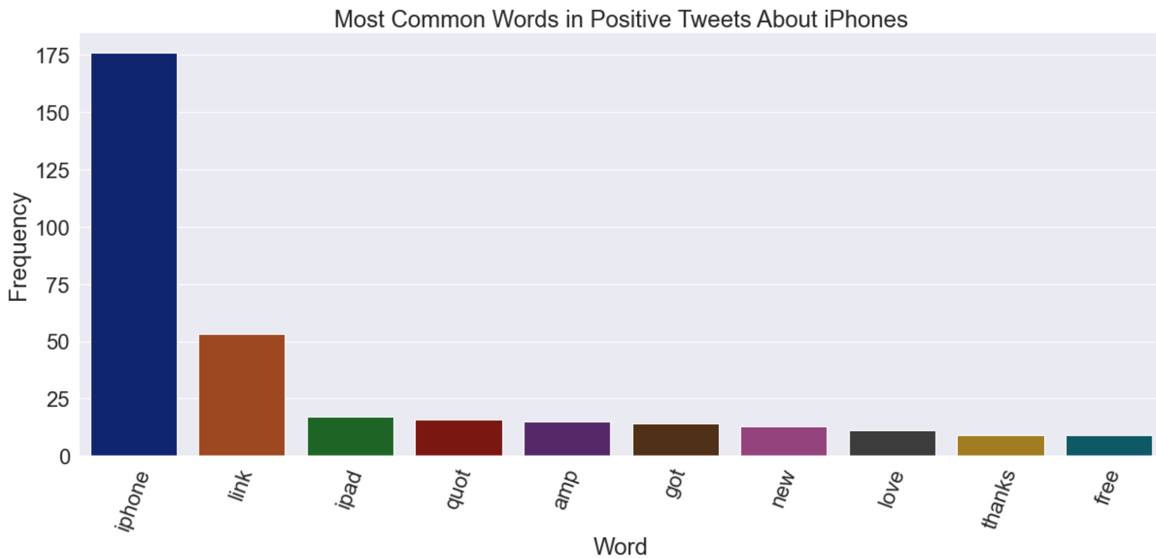
In [263]:
```
# tokenize
iphone_tokens = word_tokenize(','.join(str(v) for v in corpus_iphone))

# remove stopwords
stopped_iphone_tokens = [word.lower() for word in iphone_tokens if word.lower
                          not in stopword_list]
```

In [264]:
```
freq = FreqDist(stopped_iphone_tokens)
```

In [265]:
```
freq = pd.DataFrame(freq.most_common(25))
```

In [266]:

```
fig_dims = (20,8)
fig, ax = plt.subplots(figsize=fig_dims)
sns.set(font_scale=2)
sns.set_style("darkgrid")
palette = sns.set_palette("dark")
ax = sns.barplot(x=freq.head(10)[0], y=freq.head(10)[1], palette=palette)
ax.set(xlabel="Word",ylabel="Frequency")
plt.ticklabel_format(style='plain',axis='y')
plt.xticks(rotation=70)
plt.title('Most Common Words in Positive Tweets About iPhones')
plt.show()
```



1.11.4 Recommendation

People are happiest when their phones are charged or charging.

The positive Android Tweets are in reference to parties or people just being excited about Android phones and challenging Apple's market share. Words used include party, team, good, win, market, fan, and excited. Individual observations are not about the function of the device but rather about being a part of a new group or trend.

The positive iPhone tweets center on batteries and charging/having a charged phone to be able to use at the music festival in Austin, Texas. Words most often used include case, thanks, free, charger, wow, and best. These observations are more about being able to use iPhone to do anything, including getting funding of $130,000 in order to make a movie with only the camera on an iPhone.

In [ ]:

```
```