Identifying Negative Sentiment in Tweets with Natural Language Processing

Business Problem

Google is seeking to increase Android's share of the U.S. smartphone and tablet markets. To do so, they are seeking information on what consumers don't like about their devices. By focusing on pain points, they hope to engineer improvements that will attract and retain more customers.

While negative sentiment toward products is available in form of survey responses and customer complaints, Google also hopes to access the opinions conveyed in social media posts. To do so, they need to identify posts which express concerns and frustrations about mobile devices from among thousands of other posts. Google is frustrated that their analysts spend so much time reading through positive and neutral posts to find the negative ones, which comprise just 6% of all posts.

My task is to build a natural language processing model which can identify the negative tweets. They have asked that the model focus on catching as many negative tweets as possible, but would like analysts to be able to work at least twice as fast. So, the model should eliminate enough positive and neutral posts that negative tweets comprise at least 12% of all the tweets returned.

Data Understanding

To identify posts with gripes, I analyzed over 9,000 tweets from a dataset provided by Crowdflower via data.world. The tweets all contain references to Google or Apple products by participants in the South by Southwest (SXSW) Conference in 2011. Although the data are a decade old, and the products discussed seem ancient (e.g. iPad 2), the words used to convey negative emotions have not changed.

Each tweet in the dataset has been rated by humans as showing a positive emotion, negative emotion, or no emotion toward the Google or Apple product mentioned. A few were also labeled "I can't tell." 59% of tweets were tagged as postive, 33% as neutral, and 6% as negative, making negative tweets the smallest category aside from "I can't tell."

After tokenization, the tweets contained just 9,780 unique words, and many of these were numbers, symbols, typos, and words combined into hashtag phrases.

Data Preparation

The data contained just one null value in the tweet text, which I dropped.

To prepare the data for modeling, I combined the positive, neutral, and I-can't-tell categories in order to build a binary classification model that can identify negative tweets.

I also set aside 10% of the data as a holdout set, which I later used to validate the final model.

In [1]:

```
# import code libraries

import pandas as pd
pd.set_option('display.max_rows', 500)
pd.set_option('display.max_columns', 500)
pd.set_option('display.float_format', lambda x: '%.5f' % x)
pd.set_option('display.max_colwidth', 1000)
import numpy as np

import matplotlib.pyplot as plt
%matplotlib inline
```

```
import seaborn as sns

from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.model_selection import train_test_split, cross_val_score, KFold
from sklearn.metrics import plot_confusion_matrix, accuracy_score
from sklearn.metrics import recall_score, precision_score, f1_score
from sklearn.naive_bayes import ComplementNB, MultinomialNB

from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import string

import utils as ut
```

In [2]:

```
# import data

data = pd.read_csv('data/judge-1377884607_tweet_product_company.csv', encoding='latin-1')
```

In [3]:

```
data.head(200)
```

Out[3]:

	tweet_text	emotion_in_tweet_is_directed_at	is_there_an_emotion_directed_at_a_brand_or_produc
0	.@wesley83 I have a 3G iPhone. After 3 hrs tweeting at #RISE_Austin, it was dead! I need to upgrade. Plugin stations at #SXSW.	iPhone	Negative emotio
1	@jessedee Know about @fludapp ? Awesome iPad/iPhone app that you'll likely appreciate for its design. Also, they're giving free Ts at #SXSW	iPad or iPhone App	Positive emotio
2	@swonderlin Can not wait for #iPad 2 also. They should sale them down at #SXSW.	iPad	Positive emotio
3	@sxsw I hope this year's festival isn't as crashy as this year's iPhone app. #sxsw	iPad or iPhone App	Negative emotio
4	@sxtxstate great stuff on Fri #SXSW: Marissa Mayer (Google), Tim O'Reilly (tech books/conferences) &amp; Matt Mullenweg (Wordpress)	Google	Positive emotio
5	@teachntech00 New iPad Apps For #SpeechTherapy And Communication Are Showcased At The #SXSW Conference http://ht.ly/49n4M #iear #edchat #asd	NaN	No emotion toward brand or produc
6	NaN	NaN	No emotion toward brand or produc
7	#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	Android	Positive emotio
8	Beautifully smart and simple idea RT @madebymany @thenextweb wrote about our #hollergram iPad app for #sxsw! http://bit.ly/ieaVOB	iPad or iPhone App	Positive emotio
9	Counting down the days to #sxsw plus strong Canadian dollar means stock up on Apple gear	Apple	Positive emotio
10	Excited to meet the @samsungmobileus at #sxsw so I can show them my Sprint Galaxy S still running Android 2.1. #fail	Android	Positive emotio

running Android 2.1. #fail

	tweet_text	emotion_in_tweet_is_directed_at	is_there_an_emotion_directed_at_a_brand_or_produc
11	Find &amp; Start Impromptu Parties at #SXSW With @HurricaneParty http://bit.ly/gVLrIn I can't wait til the Android app comes out.	Android App	Positive emotio
12	Foursquare ups the game, just in time for #SXSW http://j.mp/grN7pK) - Still prefer @Gowalla by far, best looking Android app to date.	Android App	Positive emotio
13	Gotta love this #SXSW Google Calendar featuring top parties/ show cases to check out. RT @hamsandwich via @ischafer =&gt;http://bit.ly/aXZwxB	Other Google product or service	Positive emotio
14	Great #sxsw ipad app from @madebymany: http://tinyurl.com/4nqv92l	iPad or iPhone App	Positive emotio
15	haha, awesomely rad iPad app by @madebymany http://bit.ly/hTdFim #hollergram #sxsw	iPad or iPhone App	Positive emotio
16	Holler Gram for iPad on the iTunes App Store - http://t.co/kfN3f5Q (via @marc_is_ken) #sxsw	NaN	No emotion toward brand or produc
17	I just noticed DST is coming this weekend. How many iPhone users will be an hour late at SXSW come Sunday morning? #SXSW #iPhone	iPhone	Negative emotio
18	Just added my #SXSW flights to @planely. Matching people on planes/airports. Also downloaded the @KLM iPhone app, nicely done.	iPad or iPhone App	Positive emotio
19	Must have #SXSW app! RT @malbonster: Lovely review from Forbes for our SXSW iPad app Holler Gram - http://t.co/g4GZypV	iPad or iPhone App	Positive emotio
20	Need to buy an iPad2 while I'm in Austin at #sxsw. Not sure if I'll need to Q up at an Austin Apple store?	iPad	Positive emotio
21	Oh. My. God. The #SXSW app for iPad is pure, unadulterated awesome. It's easier to browse events on iPad than on the website!!!	iPad or iPhone App	Positive emotio
22	Okay, this is really it: yay new @Foursquare for #Android app!!!!11 kthxbai. #sxsw	Android App	Positive emotio
23	Photo: Just installed the #SXSW iPhone app, which is really nice! http://tumblr.com/x6t1pi6av7	iPad or iPhone App	Positive emotio
24	Really enjoying the changes in Gowalla 3.0 for Android! Looking forward to seeing what else they &amp; Foursquare have up their sleeves at #SXSW	Android App	Positive emotio
25	RT @LaurieShook: I'm looking forward to the #SMCDallas pre #SXSW party Wed., and hoping I'll win an #iPad resulting from my shameless promotion. #ChevySMC	iPad	Positive emotio
26	RT haha, awesomely rad iPad app by @madebymany http://bit.ly/hTdFim #hollergram #sxsw (via @michaelpiliero)	iPad or iPhone App	Positive emotio
27	someone started an #austin @PartnerHub group in google groups, pre-#sxsw. great idea	Other Google product or service	Positive emotio

	tweet_text	emotion_in_tweet_is_directed_at	is_there_an_emotion_directed_at_a_brand_or_produc
28	The new #4sq3 looks like it is going to rock. Update for iPhone and Android should push tonight http://bit.ly/etsbZk #SXSW #KeepAustinWeird	iPad or iPhone App	Positive emotio
29	They were right, the @gowalla 3 app on #android is sweeeeet! Nice job by the team there. #sxsw	Android App	Positive emotio
30	Very smart from @madebymany #hollergram iPad app for #sxsw! http://t.co/A3xvWc6 (may leave my vuvuzela at home now)	iPad or iPhone App	Positive emotio
31	You must have this app for your iPad if you are going to #SXSW http://itunes.apple.com/us/app/hollergram/id420666439?mt=8 #hollergram	iPad or iPhone App	Positive emotio
32	Attn: All #SXSW frineds, @mention Register for #GDGTLive and see Cobra iRadar for Android. {link}	NaN	No emotion toward brand or produc
33	Anyone at #sxsw want to sell their old iPad?	NaN	No emotion toward brand or produc
34	Anyone at #SXSW who bought the new iPad want to sell their older iPad to me?	NaN	No emotion toward brand or produc
35	At #sxsw. Oooh. RT @mention Google to Launch Major New Social Network Called Circles, Possibly Today {link}	NaN	No emotion toward brand or produc
36	The best! RT @mention Ha! First in line for #ipad2 at #sxsw &quot;pop-up&quot; Apple store was an event planner #eventprofs #pcma #engage365	iPad	Positive emotio
37	SPIN Play - a new concept in music discovery for your iPad from @mention &amp; spin.com {link} #iTunes #sxsw @mention	NaN	No emotion toward brand or produc
38	@mention - False Alarm: Google Circles Not Coming NowÛÒand Probably Not Ever? - {link} #Google #Circles #Social #SXSW	Google	Negative emotio
39	VatorNews - Google And Apple Force Print Media to Evolve? {link} #sxsw	NaN	No emotion toward brand or produc
40	@mention - Great weather to greet you for #sxsw! Still need a sweater at night..Apple putting up &quot;flash store&quot; downtown to sell iPad2	Apple	Positive emotio
41	HootSuite - HootSuite Mobile for #SXSW ~ Updates for iPhone, BlackBerry &amp; Android: Whether youÛªre getting friend... {link}	NaN	No emotion toward brand or produc
42	Hey #SXSW - How long do you think it takes us to make an iPhone case? answer @mention using #zazzlesxsw and weÛªll make you one!	NaN	No emotion toward brand or produc
43	Mashable! - The iPad 2 Takes Over SXSW [VIDEO] #ipad #sxsw #gadgets {link}	NaN	No emotion toward brand or produc
44	For I-Pad ?RT @mention New #UberSocial for #iPhone now in the App Store includes UberGuide to #SXSW sponsored by ... {link}	NaN	No emotion toward brand or produc
45	#IPad2 's Û÷#SmartCoverÛª Opens to Instant Access - I should have waited to get one! - {link} #apple #SXSW	iPad or iPhone App	Positive emotio

	tweet_text	emotion_in_tweet_is_directed_at	is_there_an_emotion_directed_at_a_brand_or_produc
46	Hand-Held Û÷HoboÛª: Drafthouse launches Û÷Hobo With a ShotgunÛª iPhone app #SXSW {link}	NaN	Positive emotio
47	HOORAY RT ÛÏ@mention Apple Is Opening A Pop-Up Store In Austin For #SXSW | @mention {link}	Apple	Positive emotio
48	Orly....? ÛÏ@mention Google set to launch new social network #Circles today at #sxswÛ	NaN	No emotion toward brand or produc
49	wooooo!!! ÛÏ@mention Apple store downtown Austin open til Midnight. #sxswÛ	Apple	Positive emotio
50	Khoi Vinh (@mention says Conde Nast's headlong rush into iPad publishing was a &quot;fundamental misunderstanding&quot; of the platform #sxsw	NaN	No emotion toward brand or produc
51	ÛÏ@mention {link} &lt;-- HELP ME FORWARD THIS DOC to all Anonymous accounts, techies,&amp; ppl who can help us JAM #libya #SXSW	NaN	No emotion toward brand or produc
52	÷¼ WHAT? ÷_ {link} ã_ #edchat #musedchat #sxsw #sxswi #classical #newTwitter	NaN	No emotion toward brand or produc
53	.@mention @mention on the location-based 'fast, fun and future' - {link} (via @mention #sxsw	NaN	No emotion toward brand or produc
54	ÛÏ@mention @mention #Google Will Connect the Digital &amp; Physical Worlds Through Mobile - {link} #sxswÛ @mention	NaN	No emotion toward brand or produc
55	ÛÏ@mention @mention talking about {link} - Google's effort to allow users to have open systems #bettercloud #sxswÛ	Google	Positive emotio
56	{link} RT @mention &quot;Google before you tweet&quot; is the new &quot;think before you speak.&quot; - Mark Belinsky, #911tweets panel at #SXSW.	NaN	No emotion toward brand or produc
57	{link} RT @mention 1st stop on the #SXSW #Chaos &amp; @mention hunt: Austin Java. Get in the spy game 4 a chance 2 win an iPad!	iPad	Positive emotio
58	{link} RT @mention Those at #SXSW check out the Holler Gram ipad app from @mention {link}	NaN	No emotion toward brand or produc
59	@mention @mention &amp; @mention having fun at #google [pic] #SXSW {link}	NaN	No emotion toward brand or produc
60	&quot;via @mention : {link} Guy Kawasaki talks 'Enchanted' at SXSW - HE knows his stuff! #books #internet #Apple #sxsw &quot;	NaN	No emotion toward brand or produc
61	#futuremf @mention {link} spec for recipes on the web, now in google search: {link} #sxsw	NaN	No emotion toward brand or produc
62	#OMFG! RT @mention Heard about Apple's pop-up store in downtown Austin? Pics are already on Gowalla: {link} #sxsw #iPad2	Apple	Positive emotio
63	#Smile RT @mention I think Apple's &quot;pop-up store&quot; in Austin would be a lot more interesting if it	Apple	No emotion toward brand or produc

actually, you know... popped up #sxsw

	tweet_text	emotion_in_tweet_is_directed_at	is_there_an_emotion_directed_at_a_brand_or_produc
64	Again? RT @mention Line at the Apple store is insane.. #sxsw	NaN	Negative emotio
65	Agree. RT @mention Wait. FIONA APPLE is in town??? Somebody kidnap her and put her in a recording studio until she records a new album. #sxsw	NaN	No emotion toward brand or produc
66	At #sxsw? @mention / @mention wanna buy you a drink. 7pm at Fado on 4th. {link} Join us!	NaN	No emotion toward brand or produc
67	attending @mention iPad design headaches #sxsw {link}	iPad	Negative emotio
68	Boooo! RT @mention Flipboard is developing an iPhone version, not Android, says @mention #sxsw	NaN	Negative emotio
69	Check out @mention @mention &amp; @mention in line for their iPad 2 in Austin. Power to them! #sxswi #SXSW {link}	iPad	Positive emotio
70	Check! RT @mention giving added value to location based services needs to battle check-in fatigue #google #pnid #sxsw	NaN	No emotion toward brand or produc
71	Chilcott: @mention #SXSW stand talking with Blogger staff. Too late to win competition for best tweet mentioning @mention So no t-shirt.	NaN	No emotion toward brand or produc
72	Do it. RT @mention Come party w/ Google tonight at #sxsw! {link} - Bands, food, art, ice cream, nifty interactive maps!	Google	Positive emotio
73	Gowalla's @mention promises to launch Foursquare check-in + Groupon rewards-type service at #SXSW. Finger's crossed. {link}	NaN	No emotion toward brand or produc
74	Ha.ha. RT @mention #SXSW News: Yahoo.com is loosing search traffic to new site, Google.com. Doubt it will last tho w/ that weird name.	NaN	No emotion toward brand or produc
75	Holla! RT @mention At google party. Best ever! Get your butt over here. #sxsw	Google	Positive emotio
76	I love my @mention iPhone case from #Sxsw but I can't get my phone out of it #fail	iPhone	Positive emotio
77	I worship @mention {link} #SXSW	NaN	No emotion toward brand or produc
78	iPad2? RT @mention Droid &amp; Mac here :) RT @mention My #agnerd confession, using laptop, iPad &amp; blackberry to follow #SXSW	NaN	No emotion toward brand or produc
79	Launching @mention #SxSW? RT @mention @mention Denies Social Network Called Circles Will Debut Today, Despite Report {link}	NaN	No emotion toward brand or produc
80	New Post: @mention iPhone app makes it easy to connect on all social networks with people you meet {link} #sxsw	iPad or iPhone App	Positive emotio
81	Nice that @mention iPhone app is behaving today. Crashes yesterday were ridiculous. #sxsw	iPad or iPhone App	Positive emotio
	Nice! RT @mention Apple opening		

	tweet_text	emotion_in_tweet_is_directed_at	is_there_an_emotion_directed_at_a_brand_or_produc
82	Nice! RT @mention Apple opening popup store for iPad [illegible] downtown Austin during #SXSW {link} via @mention	NaN	No emotion toward brand or produc
83	Nice!! RT @mention Hey, Apple fans! Get a peek at the space that's slated to be a pop-up #SXSW Apple Store tomorrow: {link}	Apple	Positive emotio
84	one thing @mention is doing so great is get a great, down to earth face to Google as a company - You can only love her #sxsw #sxwsi	Google	Positive emotio
85	Stay tune @mention showcase #H4ckers {link} #SXSW	NaN	No emotion toward brand or produc
86	Thank you @mention @mention for the #touchingstories preso #SXSW . Here's their deck {link}	NaN	No emotion toward brand or produc
87	Thank you @mention for an awesome #sxsw party! {link}	NaN	No emotion toward brand or produc
88	Thanks RT @mention If you're trying to contact friends or family in #Japan, @mention has created a person finder: {link} #SXSW	NaN	No emotion toward brand or produc
89	Thanks to @mention for her mention of our new #Speech iPad apps being showcased at the #SXSW Conf. {link} #sxswh #sxsh	iPad or iPhone App	Positive emotio
90	Thanks to @mention for publishing the news of @mention new medical Apps at the #sxswi conf. blog {link} #sxsw #sxswh	NaN	I can't te
91	Thanks to @mention for publishing the news of our new medical Apps in the #sxswi conf. blog {link} #sxsw #sxswh #mhealth	NaN	No emotion toward brand or produc
92	What !?!? @mention #SXSW does not provide iPhone chargers?!? I've changed my mind about going next year!	iPhone	Negative emotio
93	Wonder if @mention &amp; @mention will be in the apple flashmob: tcrn.ch/fcs45j #SXSW #ipad2	NaN	No emotion toward brand or produc
94	Wonder if @mention is putting tips from the @mention API... #SxSW #SUxSW	NaN	No emotion toward brand or produc
95	XMAS!! RT @mention Shiny new @mention @mention @mention apps, a new @garyvee book, pop-up iPad 2 stores... #SXSW is Christmas for nerds.	iPad	Positive emotio
96	Yai!!! RT @mention New #UberSocial for #iPhone now in the App Store includes UberGuide to #SXSW sponsored by (cont) {link}	iPhone	Positive emotio
97	Yes!!! RT @mention hey @mention , i've got another gem for you --&gt; free @mention sxsw {link} #SXSW	NaN	No emotion toward brand or produc
98	Fast, Fun &amp; Future: @mention of Google presenting at #sxsw on search, local and mobile	Google	Positive emotio
99	GSD&amp;M &amp; Google's Industry Party Tonight @mention - See u there! {link} #SXSW #Austin #Welivehere #GSDM	NaN	No emotion toward brand or produc
	New buzz? &quot;@mention Google		

	tweet_text	emotion_in_tweet_is_directed_at	is_there_an_emotion_directed_at_a_brand_or_produc
100	to Launch Major New Social Network Called Circles, Possibly Today {link} rt @mention #sxsw&quot;	NaN	No emotion toward brand or produc
101	Headline: &quot;#iPad 2 is the Must-Have Gadget at #SXSW&quot; Hmm... I could have seen that one coming! {link} #gadget	iPad	Positive emotio
102	ÛÏ@mention &quot;Apple has opened a pop-up store in Austin so the nerds in town for #SXSW can get their new iPads. {link} #wow	NaN	I can't te
103	Know that &quot;dataviz&quot; translates to &quot;satanic&quot; on an iPhone. I'm just sayin'. #sxsw	NaN	Negative emotio
104	.@mention &quot;Google launched checkins a month ago.&quot; Check ins are ok, but CHECK OUTS are the future. #sxsw #Bizzy	Google	Positive emotio
105	.@mention &quot;Google launched checkins a month ago.&quot; Check ins are ok, but CHECK OUTS are the future. #sxsw #Bizzy (via @mention	NaN	No emotion toward brand or produc
106	ÛÏ@mention &quot;Google before you tweet&quot; is the new &quot;think before you speak.&quot; - Mark Belinsky, #911tweets panel at #SXSW.Û	Google	Positive emotio
107	Attending &quot;left brain search = Google, Right brain search = X&quot; #Bettersearch -- talking about the future of search engines at #sxsw	NaN	No emotion toward brand or produc
108	#HP opens &quot;Mobile Park&quot; &amp; Content Incubator at #SXSW {link} #Apple constructs &quot;pop-up&quot; store {link}	NaN	No emotion toward brand or produc
109	Kawasaki: &quot;Not C.S. Lewis level reasoning, but Apple's continued existence is evidence for the existence of God&quot; #bawling #sxsw	Apple	Positive emotio
110	Kawasaki: &quot;pagemaker saved Apple.&quot; Oh those were the days. #sxsw #jwtatl #enchantment	NaN	No emotion toward brand or produc
111	Kawasaki: &quot;pagemaker saved Apple.&quot; Oh those were the days. #sxsw #jwtatl #enchantment via @mention	Apple	Positive emotio
112	Spark for #android is up for a #teamandroid award at #SXSW read about it here: {link}	NaN	Positive emotio
113	Unboxing. #Apple #sxsw @mention Apple Store, SXSW {link}	NaN	No emotion toward brand or produc
114	#SXSW and #Apple iPad 2's are great, but thoughts are w/ Japan and APAC regions dealing w/ earthquake &amp; tsunami trauma. #sxswi	iPad	Positive emotio
115	At #SXSW, #Apple schools the #marketing experts | SXSW - CNET Blogs {link}	NaN	No emotion toward brand or produc
116	At #SXSW, #Apple schools the marketing experts - {link}	Apple	Positive emotio
117	At #SXSW, #Apple schools the marketing experts {link}	NaN	No emotion toward brand or produc
	Temporary #apple store is def not a		

	tweet_text	emotion_in_tweet_is_directed_at	is_there_an_emotion_directed_at_a_brand_or_produc
118	tent, it's a powerhouse gym #SXSW {link}	Apple	Positive emotio
119	Temporary #Apple store on 6th and Congress for #sxsw, along with 10,000 very happy hipsters.	Apple	Positive emotio
120	ÛÏ@mention #Apple wins #SXSW {link} Opening a temporary store in downtown Austin to support #iPad2 launch - That is good.	Apple	Positive emotio
121	#iPad and #Austin are trending today. Have fun at #sxsw all you nerdy nerds!!!	iPad	Positive emotio
122	Headed to #Austin for #SXSW? Check out my map for newbies {link} @mention @mention , @mention @mention Enjoy!	NaN	No emotion toward brand or produc
123	Funny how #Austin is trending but not #SXSW. Only a matter of minutes at this point (at least according to Twitter for iPhone).	NaN	No emotion toward brand or produc
124	Christian #iPad #iPhone devs I want to talk to u at #sxsw or after -maybe we can wk together on cool app! @mention me	iPad or iPhone App	Positive emotio
125	#sxsw #ux #ipad #uxdes remember to ultimately be aware of the audience your app is targeted towards. An unexpected experience can be good.	NaN	No emotion toward brand or produc
126	The Apple #iPAD2 has taken #SxSW and #Austin by storm.. {link} @mention excited to a be a part #mobile	iPad	Positive emotio
127	I can haz #iPad2 ifrom #SxSW Gr8 {link}	iPad	Positive emotio
128	Stacks of #ipad2's waiting to be bought at #sxsw. I got mine, no hassle at all. Apple handled this perfectly {link}	iPad	Positive emotio
129	#Google's #Mobile Future, and the Elusive 'Power of Here' - {link} (via @mention #eurorscg #sxsw #sxswi	NaN	No emotion toward brand or produc
130	For those #notatSXSW (or at #SXSW), here's {link} Free to download and meet nearby peps	NaN	No emotion toward brand or produc
131	Does your #SmallBiz need reviews to play on Google Places...We got an App for that..{link} #seo #sxsw	NaN	Positive emotio
132	Does your #SmallBiz need reviews to play on Google Places...We got an App for that..{link} #seo #sxsw	NaN	No emotion toward brand or produc
133	#Samsung, #Sony follow #Apple, #HP lead @mention {link} #Austin #atx #SXSW	NaN	No emotion toward brand or produc
134	#Samsung, #Sony follow #Apple, #HP lead @mention {link} #Austin #atx #SXSW /via @mention ^rg	NaN	No emotion toward brand or produc
135	Take that #SXSW ! RT @mention Major South Korean director gets $130,000 to make a movie entirely with his iPhone. {link}	iPhone	Positive emotio
136	Beautiful #sxsw (@mention Apple Store, SXSW) [pic]: {link}	Apple	Positive emotio
137	Q1 Was at #sxsw #sxswi for prep. Amazing pre push locally. Focus on	NaN	No emotion toward brand or produc

	tweet_text	emotion_in_tweet_is_directed_at	is_there_an_emotion_directed_at_a_brand_or_produc
137	location based. Google owns 10% of the regions billboards. #pr20chat	NaN	No emotion toward brand or produc
138	Any other #Sxsw accounts I need to follow or apps to download for iPhone?	NaN	No emotion toward brand or produc
139	Headed to #sxsw and want to share/gather contact info? {link} can turn your iphone into a business card broadcaster.	NaN	No emotion toward brand or produc
140	Headed to #sxsw and want to share/gather contact info? {link} can turn your iphone into a... {link}	NaN	No emotion toward brand or produc
141	BTW - The #sxsw Apple store is sold out of all 3G models (VZW &amp; AT&amp;T).	NaN	No emotion toward brand or produc
142	Must have #SXSW app! RT @mention Lovely review from Forbes for our SXSW iPad app Holler Gram - {link}	iPad or iPhone App	Positive emotio
143	Temporary #sxsw apple store. Apple being sneaky as usual {link}	Apple	Positive emotio
144	Anyone at #sxsw been by the pop-up Apple store in Austin? That's gotta be a hopping place today.	NaN	No emotion toward brand or produc
145	ÛÏ@mention #sxsw beta testing interactive book for iPad app by Moonbot studios out of Louisiana. Cool app.Û	iPad or iPhone App	Positive emotio
146	Apple won #SxSW from day one. Seeing a TON of #iPad2	iPad	Positive emotio
147	#fastball #sxsw Giving away two NEW Ipad2 wifi 32g black Apple cover tweet @mention fo more info #sxswi #attsxsw Tonight @mention bo.lt house	NaN	No emotion toward brand or produc
148	Anyone at #sxsw had a chance to check out the pop-up Apple store? Wondering if it is worth the trek from the convention center...	NaN	No emotion toward brand or produc
149	ÛÏ@mention #sxsw ipad store sold out of everything except 64gig wifi only whiteÛ @mention Did you manage to get yours?	iPad	Positive emotio
150	ÛÏ@mention #sxsw ipad store sold out of everything except 64gig wifi only whiteÛ also known as the white jeans configuration.	iPad	Positive emotio
151	Offered a #sxsw ipad promo to ninjafinder users and fans who are not here at #sxsw. Sucks to not be here. #sxsw	iPad	Positive emotio
152	@mention #SXSW is an Austin conference. Do a google search - they have interactive / music / film.	NaN	No emotion toward brand or produc
153	Anyone at #sxsw know if apple will be (or is) selling ipad 2 there?	NaN	No emotion toward brand or produc
154	Anyone at #SXSW know if the apple store has had a new shipment of iPads yet?	NaN	No emotion toward brand or produc
155	Marc Ecko #SXSW launches #iPhone app. to autodial political change! {link} #edreform #edtech #eduVC #FightThePaddle	NaN	No emotion toward brand or produc
	#POURsite #SXSW learning about the		

	tweet_text	emotion_in_tweet_is_directed_at	is_there_an_emotion_directed_at_a_brand_or_produc
156	life-changing impact of the iPad on real people's actual lives - bravo!	iPad	Positive emotio
157	@mention #SXSW LonelyPlanet Austin guide for #iPhone is free for a limited time {link} #lp #travel	NaN	Positive emotio
158	More free #SXSW mp3 downloads, this time from iTunes: {link}	NaN	No emotion toward brand or produc
159	Anyone at #sxsw or heading to aclu event seen owt to do with google circles then?	NaN	No emotion toward brand or produc
160	@mention #SXSW prompt for memory: go to Google map and describe a childhood walk	NaN	No emotion toward brand or produc
161	The Apple #SXSW store still has iPad 2's and short lines.	iPad	Positive emotio
162	Essential #sxsw tools: {link}	NaN	No emotion toward brand or produc
163	Just left #sxsw tradeshow demo of @mention at the Google Theatre. Ok, I get it. I see why all the presenters here are using it.	Other Google product or service	Positive emotio
164	Following #sxsw Tweets on Google Realtime, four platforms on Tweet Deck and listening to panel, realizing I'm spoken to no one here today.	NaN	No emotion toward brand or produc
165	Anyone at #sxsw want an iPad 2? I'm in line and will pick one up for someone willing to pay me 50 for me to grab 1 for you?	NaN	No emotion toward brand or produc
166	Anyone at #sxsw want to make a quick hundred dollars? New #ipad2 from ad hoc apple store here gets hundred plus cost!	iPad	Positive emotio
167	Solving a #SXSW-induced iPhone-in-toilet crisis at Apple Store with @mention (not my crisis for once)	NaN	No emotion toward brand or produc
168	Monday at #sxsw: barry diller, new york times, congress lunch, W Hotel party, Google party, six dirty martinis. How mondays should be.	Google	Positive emotio
169	Attending #sxsw? Austin Guide by @mention is now free to download on iTunes - {link} #lp	NaN	No emotion toward brand or produc
170	Seriously #sxsw? Did you do any testing on the mobile apps? Constant iPad crashes causing lost schedules, and no sync for WP7.	iPad or iPhone App	Negative emotio
171	Ready for #SXSW?! Here are some iPhone apps 2 make ur blogging easier | {link} #SXSW #SXSWi	iPad or iPhone App	Positive emotio
172	ipad2 and #sxsw...a conflagration of doofusness. {link}	iPad	Negative emotio
173	attention #sxsw'ers - {link} - rumored pop-up temporary Apple store for all your #sxsw iPad 2 launch needs.	Apple	Positive emotio
174	I went to #sxswi and all I won was this lousy #iPad #sxsw :-) :-) {link}	iPad	Positive emotio
175	Hey #sxsw #sxswi folks. If you want to learn about security come over to #bsidesaustin {link}	NaN	No emotion toward brand or produc
176	Attending #SXSWi? Work in iPhone / iPad game development? Looking to hire an Austin-based iOS developer? I'm your man. Let's talk. #SXSW	NaN	No emotion toward brand or produc

	tweet_text	emotion_in_tweet_is_directed_at	is_there_an_emotion_directed_at_a_brand_or_produc
177	This is a #WINNING picture #android #google #sxsw {link}	Android	Positive emotio
178	GSD&amp;M + Google 7-10. RT @mention What's the best party to hit tonight? #sxsw @mention @mention	NaN	No emotion toward brand or produc
179	GSD&amp;M + Google Industry Party #SXSW @mention great to meet you {link}	NaN	No emotion toward brand or produc
180	You spent $1,000+ to come to SXSW. \n\nYou've already used iPad 1. \n\nThe wait is a couple city blocks. \n\nWhy? #ipad2 #SXSW {link}	iPad	Negative emotio
181	#sxsw day 1 - Marissa Mayer: Google Will Connect the Digital &amp; Physical Worlds Through Mobile {link}	NaN	No emotion toward brand or produc
182	Behind on 100s of emails? Give them all 1 line iPhone composed replies. #SXSW #protip	iPhone	Positive emotio
183	It's like 10pm at night and there is a line around the block at the popup apple stores selling iPad2s. #sxsw	iPad	Positive emotio
184	.@mention 1154 free songs from #SXSW (this year alone!) {link}	NaN	No emotion toward brand or produc
185	more than 150 million mobile users for Google Maps for Mobile #SXSW	NaN	No emotion toward brand or produc
186	Currently 150 people in line at the &quot;Pop Up Apple Store&quot; #sxsw	NaN	No emotion toward brand or produc
187	Only iPad 2 available at #sxsw is the 64GB wifi-only model at 699, plustheoptional(not69 !)leathersmartcoverat	NaN	No emotion toward brand or produc
188	÷¼ We love 2 entertain youÛ_Please donÛªt be grateful! ÷_ {link} ã_ #edchat #musedchat #sxsw #sxswi #classical #newTwitter	NaN	No emotion toward brand or produc
189	Less than 2 hours until we announce the details on the iPad 2 giveaway! #SXSW #SXSWi	NaN	No emotion toward brand or produc
190	I'm up to 2 iPad 2s seen in the wild. Both people say it is fast, but the still pics are terrible. #sxsw	iPad	Negative emotio
191	(The iPad 2 queue at #sxsw of course	NaN	No emotion toward brand or produc
192	The #iPad 2 Takes Over #SXSW [VIDEO] /by @mention {link}	iPad	Positive emotio
193	many iPad 2's snapping away at the keynote slides! #sxsw	iPad	Positive emotio
194	Apple has 200M users' credit cards sync'd with iTunes for one click purchase. . #winning #sxsw	NaN	No emotion toward brand or produc
195	Having my 2nd cocktail &quot;Texas Snowflake&quot; (@mention google it!!) at the #CNNGrill #SxSw	NaN	No emotion toward brand or produc
196	New Post: 3 iPhone Apps We'll Be Using at South By Southwest Interactive {link} #SXSW #SXSWi	iPad or iPhone App	Positive emotio
197	New Post: 3 iPhone Apps We'll Be Using at South By Southwest Interactive #SXSW #SXSWi {link}	NaN	No emotion toward brand or produc
198	sweet new 3-d google maps demo going on in ballroom D #SXSW	Other Google product or service	Positive emotio

	tweet_text	emotion_in_tweet_is_directed_at	is_there_an_emotion_directed_at_a_brand_or_produc
199	More than 35 million miles per day are driving with Google Maps Navigation! #SXSW	Other Google product or service	Positive emotioi

In [4]:

```
# check data length

len(data)
```

Out[4]:

```
9093
```

In [5]:

```
# change column names to 'text' and 'label'

data.rename(columns={'tweet_text':'text', 'is_there_an_emotion_directed_at_a_brand_or_pro
duct': 'label'}, inplace=True)
```

In [6]:

```
# drop product column since this model will only predict sentiment, not the product as we
ll

data.drop(columns = 'emotion_in_tweet_is_directed_at', inplace=True)
```

In [7]:

```
# data has one null value in the text

data.isna().sum()
```

Out[7]:

```
text      1
label     0
dtype: int64
```

In [8]:

```
# drop null value

data.dropna(inplace=True)
```

In [9]:

```
# recheck data length

len(data)
```

Out[9]:

```
9092
```

In [10]:

```
# reset index after dropping null value so train/test splits work later on

data.reset_index(drop=True, inplace=True)
```

In [11]:

```
# check label distribution

data['label'].value_counts(normalize=True)
```

Out[11]:

```
No emotion toward brand or product    0.59261
```

```
Positive emotion                      0.32754
Negative emotion                      0.06269
I can't tell                          0.01716
Name: label, dtype: float64
```

In [12]:

```
# create temporary df to plot label distribution

df_plot = pd.DataFrame(data['label'].value_counts(normalize=True)).reset_index()
df_plot.rename(columns={'label': 'Per Cent of Tweets', 'index': 'Sentiment'}, inplace=True)
df_plot['Sentiment'] = df_plot['Sentiment'].map(lambda x: 'No emotion' if x == 'No emotion toward brand or product'
                                                 else x)
df_plot['Per Cent of Tweets'] = df_plot['Per Cent of Tweets'].map(lambda x: round(x*100, 2))
```

In [13]:

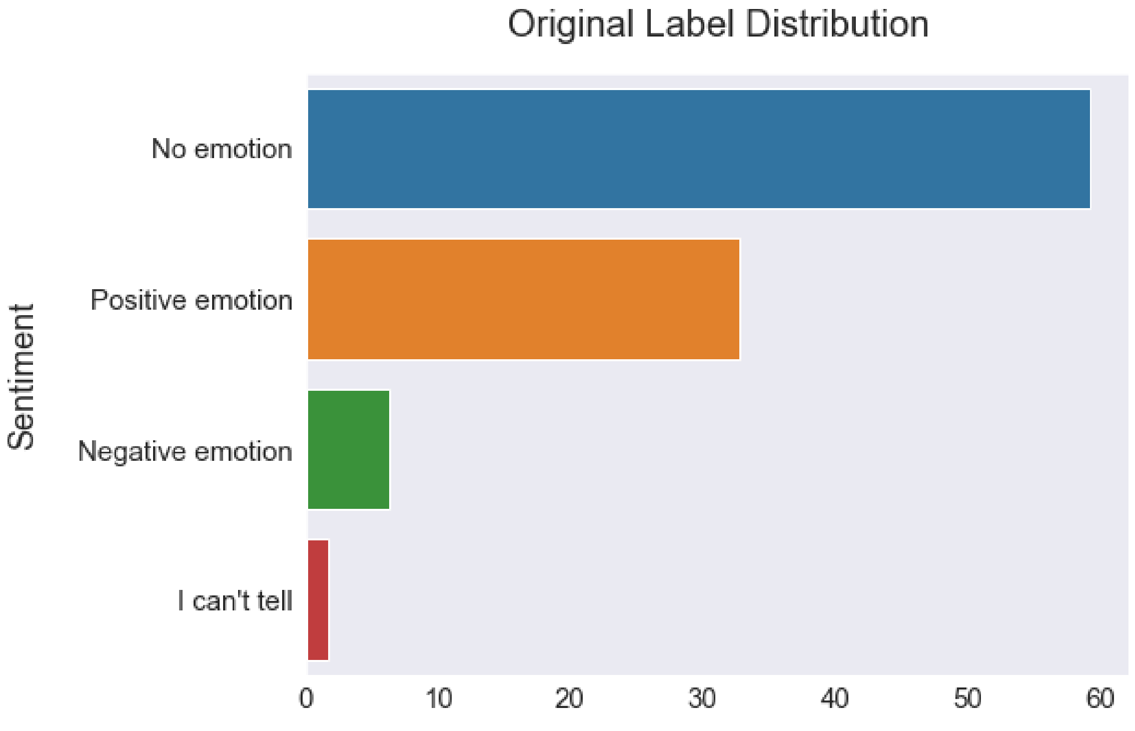
```
df_plot
```

Out[13]:

	Sentiment	Per Cent of Tweets
0	No emotion	59.26000
1	Positive emotion	32.75000
2	Negative emotion	6.27000
3	I can't tell	1.72000

In [14]:

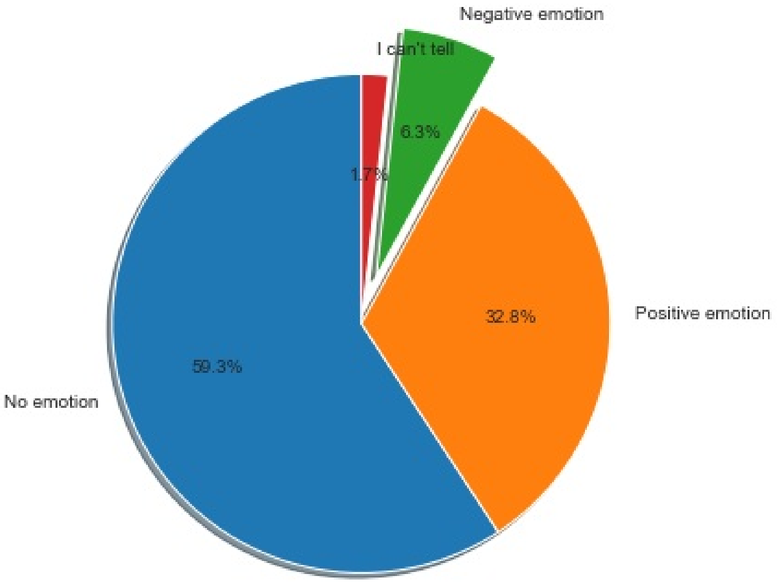
```
# plot label distribution

fig = plt.figure(figsize=(8,6))
sns.set_style('dark')
sns.barplot(x='Per Cent of Tweets', y='Sentiment', data=df_plot, orient='h')
plt.title('Original Label Distribution', fontsize=20, pad=20)
plt.xlabel('Per Cent of Tweets', fontsize=18, labelpad=20)
plt.ylabel('Sentiment', fontsize=18, labelpad=20)
plt.xticks(fontsize = 15)
plt.yticks(fontsize = 15);
plt.savefig('images/orig-label-distribution', bbox_inches='tight')
```



Per Cent of Tweets

In [15]:

```
# plot label dist in pie chart

labels = df_plot['Sentiment']
sizes = df_plot['Per Cent of Tweets']
explode = (0, 0, 0.2, 0)

fig1, ax1 = plt.subplots(figsize=(8,6))
ax1.pie(sizes, explode=explode, labels=labels, autopct='%1.1f%%',
        shadow=True, startangle=90, )
ax1.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.

plt.show()
```



In [16]:

```
# combine "No emotion toward brand or product", "I can't tell", and "Positive emotion"
# since the goal is to find negative tweets

data['label'] = data['label'].map(lambda x: 'Not negative' if x != "Negative emotion"
                                 else x)
```

In [17]:

```
# check out new label distribution

data['label'].value_counts(normalize=True)
```

Out[17]:

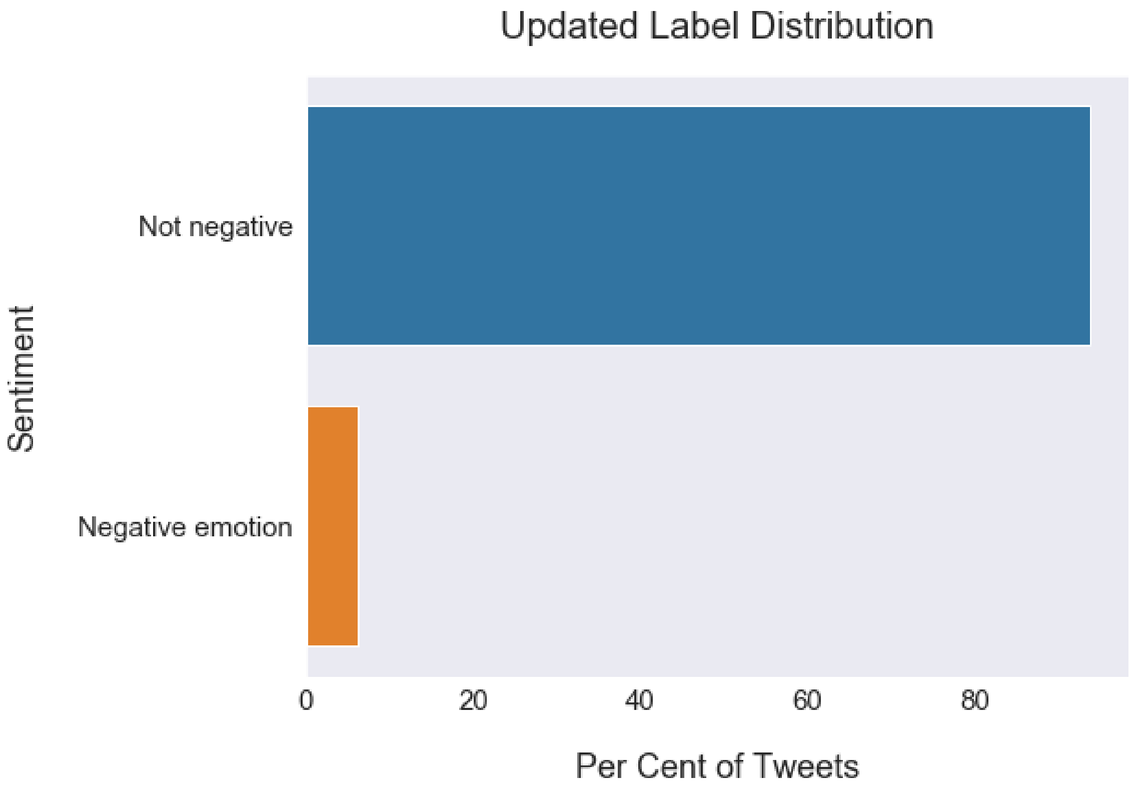
```
Not negative        0.93731
Negative emotion    0.06269
Name: label, dtype: float64
```

In [18]:

```
# create temporary df to plot new label distribution

df_plot_2 = pd.DataFrame(data['label'].value_counts(normalize=True)).reset_index()
df_plot_2.rename(columns={'label': 'Per Cent of Tweets', 'index': 'Sentiment'}, inplace=
True)
df_plot_2['Sentiment'] = df_plot_2['Sentiment'].map(lambda x: 'No emotion' if x == 'No e
motion toward brand or product'
                                               else x)
df_plot_2['Per Cent of Tweets'] = df_plot_2['Per Cent of Tweets'].map(lambda x: round(x*1
00, 2))
```

In [19]:

```
# plot new label distribution

fig = plt.figure(figsize=(8,6))
sns.set_style('dark')
sns.barplot(x='Per Cent of Tweets', y='Sentiment', data=df_plot_2, orient='h')
plt.title('Updated Label Distribution', fontsize=20, pad=20)
plt.xlabel('Per Cent of Tweets', fontsize=18, labelpad=20)
plt.ylabel('Sentiment', fontsize=18, labelpad=20)
plt.xticks(fontsize = 15)
plt.yticks(fontsize = 15);
plt.savefig('images/upd-label-distribution', bbox_inches='tight')
```



In [20]:

```
# create X and y

X = data['text']
y = data['label']
```

In [21]:

```
# reformat X to lowercase and string (some tweets were numeric)

X = X.astype(str).map(lambda x: x.lower())
```

In [22]:

```
# create holdout set as 10% of data.  Set aside until final model validation

X_train, X_holdout, y_train, y_holdout = train_test_split(X, y, test_size=0.1, random_st
ate=9117)

print(len(X_train), len(X_holdout))
```

```
8182 910
```

In [23]:

```
# cast X_train as string to make sure no values are numeric

X_train = X_train.astype(str).map(lambda x: x.lower())
```

In [24]:

In [24]:

```
X_train.tail(200)
```

Out[24]:

```
640                                 i knew if i plied @mention with beer and stogies las
t night i'd weasel my way into the team android party tonight. #success #sxsw.
135
beautiful #sxsw (@mention apple store, sxsw) [pic]: {link}
888                                                                     hootsuite mobil
e for #sxsw ~ updates for iphone, blackberry &amp; android {link} (via @mention
7784                                           @mention rt: new #ubersocial for #iphone
now in the app store includes uberguide to #sxsw sponsored by #mashable {link}
851                                            for those looking for hig-like guidelines
when designing for android {link} by adam beckley #uxamandroid @mention #sxsw
195                                                                       new post: 3 i
phone apps we'll be using at south by southwest interactive {link} #sxsw #sxswi
4150                                        and how clever that the ipad2 comes out the
first day of #sxsw with a pop up apple store right indoor town austin. clever!
4329                      fab! rt @mention rt @mention so @mention just spilled the beans
: next platform 4 #flipboard is the iphone.workin on it. #sxflip #sxsw #sxswi
2832
just joined the heaving river flow into see marissa mayer (google) #sxsw
3635                                                                 @mention if you're
looking for a space to #escape at #sxsw, why don't you stop by arthaus! {link}
691                             the first ipad didn't even exist here last year and i alr
eady feel like i'm pulling out an antique everytime i use my ipad #sxsw #ipad2
3342                                                                     @mention googl
e plze tammi.  i'm in middle of #sxsw craziness and everything is soooooo busy!
2020
mom...i want my ipad 2 back! {link} #thingsthatdontgotogether #sxsw lisa rinna
7659
lots o' free music from #sxsw bands: {link} and {link}
1127                                                                                 is
it bad that i just want to go home and read my google reader feed? #geek #sxsw
8240                                                                             if i d
on't have my iphone back by #sxsw idk what i'll do..follow the masses i guess..
6977                             rt @mention yayrt @mention new #ubersocial for #iphone
now in the app store includes uberguide to #sxsw sponsored by #mashable {link}
4460                              having fun w/ @mention new check-in's feature on iphon
e | see @mention latest article &quot;roll your own 4square&quot; {link} #sxsw
7475                                                                           how to i
mprove website rankings: advice from google and bing at #sxsw | poynter. {link}
362                             texas has been amazing i've met so many influential peop
le that work at twitter, foursquare, microsoft and even apple.  #sxsw #winning
257
did u see anything on google's circles at the #sxsw? @mention @mention
4006
heartbreaker #sxsw #apple #ipad2 rt @mention @mention just asked. sadly no :(
6293                                                    rt @mention marissa mayer: googl
e will connect the digital &amp; physical worlds through mobile - {link} #sxsw
2204                                           marissa mayer @mention : google will conn
ect the digital &amp; physical worlds through mobile {link} #sxsw via @mention
1693                        @mention #qagb with @mention listening to @mention and @menti
on talk about website ranking with google and bing #sxswi #sxsw #google #bing
3819
i won an ipad at #sxsw! nah i'm lying, i bought this shit myself :( {link}
4472
no, i didn't get an ipad 2 :( no, i'm not at #sxsw :( yes, i am depressed :(
4296                                                 nuts.  ûï@mention @mention (via @me
ntion #sxsw ipad store sold out of everything except 64gig wifi only whiteû
2874                       #sxsw panel: &quot;staying alive: can indie iphone game devel
opment survive?&quot; kind of a downer... they should try #coronasdk! @mention
748
google no lanzara ningun producto en south by southwest #sxsw 2011 {link}
8669                                                                                  g
oogle launching social network &quot;google circles&quot; @mention  #sxsw ?????
7029                                         so many good places in here rt @mention if
you're racing around #sxsw you best be fueling up with great local fare {link}
579                                    ûï@mention google to launch major new social netwo
rk called circles, possibly today {link} #sxswû sta, nije im dosta gbuzz-a?
7229
```

```
checking out @mention - iphone app for finding a car service. #sxsw
338
if ur not at the #google #aclu 80's party....u should be! #sxsw
5356                                                      rt @mention 4g will do for con
nectivity what the iphone did for smart phones - joe berry #sxsw #connectedcar
4586
get your new wordpress blog indexed in 24 hours [checklist] {link} #sxsw
8117
google tests ûïcheck-in offersû at #sxsw {link}
8231                                                                             the fl
ight from sf to austin is filled with google tshirts and youtube fleeces. #sxsw
7266                                                            curious how ipad 2 sales
went in austin, tx where a lot of potential ipad 2 buyers are attending #sxsw
1016                  brilliant pr stunt, business idea and customer service response in
one: apple sets up a pop-up store at #sxsw, draws crowds, media &amp; revenue
2229                                                                marissa mayer: goog
le will connect the digital &amp; physical worlds through mobile - {link} #sxsw
5052               rt @mention .@mention of @mention on manufacturing serendipity- &quot;
having access to more information makes us more curious, not less&quot; #sxsw
8450                                                               n26: set the terror
level to red {link} [codes valid: 4:00-7:59:59p 03/13/11] #infektd #sxsw #necro
3296                               #sxsw google party: league of extraordinary h4ackers
promoting txt redcross to 90999 to help japan @mention speakeasy {link} #photo
7863                                                                      ipad users ha
ve slower and more leisurely usage than iphone users. @mention #tapworthy #sxsw
8990                                         bing party in the same location as last yea
r's google party #irony #sxsw (@mention six lounge w/ @mention @mention {link}
6396
rt @mention official #sxsw app û÷sxsw goûª {link} #android #iphone #ipad
7846                                                                             ipad 2
has been purchased. it's 90 and sunny.  tan. 4square. so far so good at #sxsw.
1594                                        nice. rt @mention hey, apple fans! get a pee
k at the space that's slated to be a pop-up #sxsw apple store tomorrow: {link}
5029                                                                   rt @mention    g
oogle (tries again) to launch a new social network called circles: {link} #sxsw
6709                                                                         rt @mentio
n temporary #apple store is def not a tent, it's a powerhouse gym #sxsw  {link}
4154                       my sis julie and i are in a life and death rockaroke struggle
to win her an ipad at the #fandango #sxsw party. hole and bad religion so far
5246                          rt @mention #qagb #sxsw timely! rt @mention bing's search
engine share continues to rise, up to 13.6%. google still tops at 65.4% {link}
2418                                                                    #bjdproductions
#lightbox_photos wants to be your new #android camera app (#sxsw) {link} #tech
1951                                                                      @mention also
if your at #sxsw and have an iphone you can use the @mention site to check in.
3328
my #sxsw google calendar is getting a little out of control
8491                                          who's sitting in the lobby of her hotel af
ter 2am for free wifi so she can set up her new ipad?  yeah, that's me.  #sxsw
4636                                                     interesting. rt @mention rt @me
ntion google circles might launch today at #sxsw, a new social network: {link}
6063                         rt @mention i think my effing hubby is in line for an #ipad
2. can someone point him towards the line-up for wife number #2. #sxswi #sxsw
2895                      apple2 open 'pop up' temporary store @mention #sxsw 4 ipad 2 l
aunch: apple2 open 'pop up' temporary store @mention sxsw 4 ipad 2 l #: {link}
3988                                                                           c23: che
ck the head  {link} [codes valid: 12:00-3:59:59p 03/13/11] #infektd #sxsw #cvdc
6891                       rt @mention we're co-hosting a cmty mngr meetup w/ @mention @m
ention @mention at etsy austin space on 6th. 6-8p. google schwag. come! #sxsw
7750                                        @mention rt @mention hoot! new blog post: ho
otsuite mobile for #sxsw ~ updates for iphone, blackberry &amp; android {link}
8229                                                             some great free music!
-- \n20+ free tracks- #sxsw music sampler available on @mention @mention ) #fb
7116
#apple #popupstore #sxsw.  get your #ipad here  0310apple {link}
7288                                                the linkdown (non-#sxsw edition): ip
ad art show, rise austin, world poopin' day, social media events, more: {link}
3145
@mention feature @mention for the iphone. we will be all over #sxsw this week!
7133    nyt app for ipad: not &quot;here's an amazing way to serve our readership,&quot;
more &quot;here's a market opportunity we can't ignore.&quot; #sxsw #newsapps
7459                                                            all lbs apps on my iphon
```

```
e think i am in orlando :)) ! i guess i have 2 check in at disney vs #sxsw :))
3446                                                                                  d
eep in the heart of texas...with my ipad and a margarita in hand...#sxsw {link}
1134                                                              pre-order the @mention
'humorous to bees' lp on @mention catch them this month @mention {link} #sxsw
3087                         youtube gets 2b views/day, 10% are mobile. 35 hours of video
uploaded every minute. @mention google league of extraordinary hackers #sxsw
4505
gee, i wonder why so many of my android apps are updating... #sxsw
6011                                              rt @mention hey, apple fans! get a pee
k at the space that's slated to be a pop-up #sxsw apple store tomorrow: {link}
7303                                                                        n3: a slap
in the face {link} [codes valid: 12:00-3:59:59p 03/11/11] #infektd #sxsw #necro
583                                        ûï@mention google to launch major new social
network called circles, possibly today at sxsw: {link} #sxsw #nfusionû #hmm
6261                                                                           rt @ment
ion location, location location ! {link} via @mention from #google during #sxsw
4557                                                                       get in line!
things will probably get crazy at the temp #apple store in a bit. {link} #sxsw
2169                            we're creatures of habit. google found that ctrs for blu
e links far outpaced green. the darker the link, the more clicks it got. #sxsw
7438                                                                     planzai app fe
aturing the official british music schedule now on android market! {link} #sxsw
5146                            rt @mention @mention rainey street is a nice alternative
to 6th street #sxsw parties. walking directions from convention center {link}
8742
my ipad auto completes kawasaki's name from the first four letters #sxsw
8684                                                                     i messed up an
d didn't bring iphone charger. anyone got one i could borrow for 10 mins? #sxsw
6077               rt @mention i'm debuting my new iphone &amp; droid app at #sxsw next y
ear. it makes your phone waterproof &amp; will light your cigarettes for you.
3394
{link} ohh marrisa, people are hating on you, huh? #google #sxsw
6337                         rt @mention mypov: winner: popup apple store, chevy cruze
losers: investors propping up frothy startups w/ no enterprise strategy. #sxsw
2816                                                                               if y
ou need access to atx hackerspace during #sxsw, hit up our google group: {link}
3641                                                                            and the
biggest line at #sxsw?  to buy an ipad 2.  every other compan has an open bar.
2864                                                                        google pref
ers to launch hyped new social features with meh, not bang? via tc {link} #sxsw
2029                                            great recap of a seemingly great #sxsw s
ession: rt @mention relive the wonder that was the google v bing panel: {link}
2640                                             guy on the couch playing with his white
ipad 2 while mashbash parties around him is making a statement, dammit. #sxsw
2100                                                        google's art project would b
e a great virtual field trip for kids! museums around the world! #edtech #sxsw
3324                                                              @mention good to know,
and smart. sure many people who will need apple producs t.{link} #sxsw #sxswi
1758                                                              i'll pay $900 for a ne
w ipad 2, white, 32 gb, 3g in the next 20 hours. {link} #willpay #sxsw #zaarly
3219                                     #mullenweg admits that iphone app for wordpress is
not very good yet. which is very true. respect his honesty and awareness #sxsw
2828                                                    just showed off @mention charge anyw
here at a bar to charge my iphone and the whole table wanted to buy one. #sxsw
1236                                                                                app
le set to open popup shop in core of #sxsw action {link} (ipad2 on the ground.)
6766                                      rt @mention the session #designingforkids is chang
ing my mind about my future kid's relationship with the iphone. #sapient #sxsw
3223                                       so the big buzz this year at #sxsw, ipad 2, of co
urse, and group chat/text services like group me or yobongo. now you know. #fb
229
apple iie ad in the '85 si swimsuit issue at a garage sale #sxsw {link}
5752                                        rt @mention front gate tickets present the morni
ng after party 3/18 https://sites.google.com/site/frontgatesxsw11/ #sxsw music
6813                 rt @mention tonight, @mention is checking out the kills and @mention
is checking out the gsd&amp;m/google party. come say hi. #sxsw #sxswi #music
6786                                                             rt @mention this google
/bing q&amp;a panel is like the world's most expensive seo consultation. #sxsw
6637                                                      rt @mention rumors of an apple
store opening for #sxsw at 6th &amp; congress. all signs point to yes! {link}
7916                                                               #companies to watch,
from the #sxsw trade show floor {link} #apps #features #hardware #ipad #iphone
618                                                 @mention  hello! enjoy #sxsw and ri
de anywhere in austin for $10 . download the #groundlink app, {link} booth 437
2345
@mention check out @mention awesome dj skills on the ipad while u are at #sxsw
6529                                                              rt @mention rt @menti
on free itunes album, #sxsw featured artists, grab it if you missed it: {link}
900                                                 having so much fun handing out chan
ces to win 2 audi cars with @mention #sxsw  (@mention apple store sxsw) {link}
6112                                         rt @mention if you're in austin today, plea
se join @mention @mention and myself talking ipad 2 and #sxsw  here: {link} :)
5624                           rt @mention co founder of google teacher academy is bringi
ng hs students from austin to show how they use tech: {link} #sxsw #education
1928                             existential google. mt @mention ûïcontextual discoveryû
demo: where you are, your history, time of search to refine results. #sxsw
6907                                      rt @mention we're not launching any products a
t #sxsw we're doing plenty else. join us for #h4ckers &amp; 80s dancing {link}
8047                             google's social network launch? parcelgenie.com at sxsw
i and hears rumours that google's 'google circles' will launch today ... #sxsw
3240                                                              refreshing, liveley
and informative talk on game design mechanics by @mention @mention #sxswi #sxsw
8138
oh at #sxsw. oh, that's just an ipad 1, i thought he was #winning.
2114
is the flash discussion still relevant? #sxsw #tapworthy ipad design headaches
1943                                                             #virtualwallet #sxsw
no nfc in #iphone5 bc of standardization while #android will have it #confusion
4506                                                              bereft wanderer. whi
te cord, limp. lifeless. there is no outlet for your iphone here. #sxsw #poetry
6970                                                                      rt @mention
woot!! just won the #google #lego hackathon competition. #sxsw #startupbus #cle
220                         just took a survey on iphone while in starbucks line. got a
free starbucks gift card. instant research. instant gratification. #sxsw #gsdm
3329                                                                           @mentio
n google circles by @mention stresses context, not sharing with everyone. #sxsw
2375
pretty excited for my iphone to stop working #sxsw #at&amp;t
335                                                              #technews at sxsw, app
le schools the marketing experts {link} #tech_news #apple #jobs_co #sxsw #tech
4035                                              it is well known steve jobs hates #sxs
w rt @mention iphone just autocorrected &quot;sxsw&quot; to &quot;ass's&quot;.
1980                   just saw an iphone #periscope move w/ 2nd camera @mention  gowalla
talk @mention #sxsw. snoring guy busted &amp; documented. get a cpap, homie!
2705                                             use google profile or fb as  entry poin
t? fb too personal? try digg or google reader to draw people in. #hireme #sxsw
789                                                                       google to la
unch major new social network called circles, possibly today {link} #sxsw&quot;
4757                            #sxsw @mention #devops: @mention (umm. meant, go google
&quot;why complex systems fail&quot;, written by doctor. reads like it guy. :)
6029                                                            rt @mention hoot new blo
g post:hootsuite mobile for #sxsw~updates for iphone,blackberry,android {link}
4429                                           i'd pay an ipad 2 to the person who gets
the most zaarly referrals by march 12th. #zaarlyiscoming #winning #sxsw {link}
7486                                                  google maps mobile route around tr
affic feature reduces fuel consumption and time spent in traffic. {link} #sxsw
1322                  companies who are embracing nfc today: google (nfc window decal &am
p; nexus s), nokia (willinclude in all smartphones  in 2011) #sxsw #mcommerce
3569                                             @mention i was there ~5:30 and the line
was around the block. decided to forgo ipad 2 goodness for food. #sxsw #apple
4165
apple declines to be at the html5 browser wars iv panel #sxsw
2237                         marissa mayer: location and contextual discovery will enabl
e mobile devices to make us more efficient.  e.g. google places w/hotpot #sxsw
8555                                                                             #gowa
lla to launch &quot;groupon or living social-type&quot; rewards at #sxsw {link}
1053         i like it rt @mention @mention &quot;google before you tweet&quot; is the ne
w &quot;think before you speak.&quot; - mark belinsky #911tweets panel #sxsw.
4848                                                   i guess no google social network
just yet. but soon? #socialmedia #sxsw #facebook #monopoly #integration {link}
4489                       james franco is going over notes for his #sxsw speech ipad 2
vs android vs world. he got an ipad days ago but feels like an expert already.
2916                                                             #apple opening pop-up
```

```
store in austin for #sxsw geekfest [apple] {link} #applesxsw #southbysouthwest
2590                                    srsly love @mention @mention promo @mention srsly
hate that it excludes @mention esp. since my ipad insists i'm at disney #sxsw
5669                                                   rt @mention deviantart buys 3 ipad
2's in austin, tests muro drawing, it's super fast!! #deviantart #sxsw {link}
6255                                                                rt @mention line is w
rapping around the block for an ipad 2 again for a second day at #sxsw! {link}
8666
google launching new social network called cicles at #sxsw today {link}
1815                                            just got a free iphone-charger from alex
on the #powermatteam. wow. makes life at #sxsw so much easier! thanks!  {link}
1149
phrase of the day iphone ready android coming #sxsw
4744                       always wanted this! rt @mention sound of my voice was shot expl
oiting apple &amp; best buy's 14-day return policy on imacs. brilliant. #sxsw
2043
queue at apple pop-up store at #sxsw still long!
8592                                                                              google int
roducing check-in status and rewards at #sxsw - mobile is where it's at. {link}
2363
google social network: circles   coming today at #sxsw? {link}
7232                                                               checking out ipad 2 @m
ention the #sxsw pop-up store (@mention apple store, sxsw w/ 17 others) {link}
571                                     ûï@mention google to launch major new social netwo
rk called circles, possibly today {link} #sxswû \n\nsomething to keep watch
5378                                          rt @mention akqa is hiring. find me up front
after designing ipad interfaces - new navigation schemas in ballroom a. #sxsw
1845                                               there is a tech bro posed as a homeles
s person outside the apple pop up with a sign asking for $$$ for an ipad #sxsw
5118                              rt @mention @mention cool! that means we can watch ustr
eam in skyfire browser on iphone :-)\n(safari doesn't work) \nhave fun!  #sxsw
8641                                        how do you use maps? mayer: 40% of google map
s usage is mobile (there r 150 million mobile users) {link} via @mention #sxsw
358                                                      this will be fun to watch. #ipad
madness rt @mention apple opening pop-up store in austin for sxsw {link} #sxsw
2448                                                major ipad design flaw: the sxsw go i
pad app. it doesn't stay open when you switch apps! #ipaddesignheadaches #sxsw
1876                                                           win an ipad 2 from @mentio
n - just submit &amp; vote up your favorite quotes from #sxsw at {link} {link}
7469                                               ze frank project: walk down google str
eetview down a street u've walked many times b4, and revelations pop up. #sxsw
3107                                                                           z28: curf
ew be damned  {link} [codes valid: 8:00-11:59:59p 03/13/11] #infektd #sxsw #zlf
783                                                                  google to launch ma
jor new social network called circles, possibly today {link} #sxsw via @mention
1679                                               @mention #bt #sxsw &quot;having a real
ly great social search is probably a very good idea for #google&quot; @mention
2273                                                                      @mention blogg
ing from your ipad notes the next step. hope you are having a great time. #sxsw
3533                                  just bought one of the last few ipads at the apple
store in downtown austin at #sxsw. they sold out before i was done purchasing.
1103                                         i need to start downloading more apps on my
ipad and play with them for inspiration. what are your favorites? #uxdes #sxsw
19                                                           need to buy an ipad2 while i
'm in austin at #sxsw. not sure if i'll need to q up at an austin apple store?
55                    {link} rt @mention &quot;google before you tweet&quot; is the new &q
uot;think before you speak.&quot; - mark belinsky, #911tweets panel at #sxsw.
784                                                         google to launch major new so
cial network called circles, possibly today {link} #sxsw via @mention @mention
6035
rt @mention hotpot #google #marissameyer what is next #clevelandsteamer. #sxsw
1126                               for those that can't wait. 6th and congress get ur ipa
d 2. rt @mention apple opening pop-up store in austin for #sxsw tcrn.ch/eb5fjs
5528             rt @mention audience q: what prototyping tools do you use? sketchbooks/sh
arpie pens, photoshop, balsamic, google docs, axsure, etc. #myprototype #sxsw
1530                                    (cnnmoney) for #sxsw 2011, any computing device b
igger than an ipad is passì©. the mobile space has all the buzz {link} #wssxsw
2942
#sxsw  @mention designing ipad interfaces - new navigation schemas {link}
1073                                                expect to see several nfc  trials thi
s yr, google &amp; android working w/ @mention #virtualwallet #digitalid #sxsw
4113
```

```
#sxsw apple store run out for the day :( boo apple.
7202                                                  glad i brought my #mac to #sxsw! p
c clearly not cool in this environment. may walk around with #theplatform ipad
1138                                                                                  p
ics from the #apple #ipad2 line at #sxsw #fb  {link} {link} http://t.co/26svo3m
4014
awaiting keynote speaker chris poole. #sxsw #iphone  #twitpict {link}
1691                                                                                  s
cored a #mophie juice pack at the #tradeshow #sxsw. double your iphone battery!
3884                                    #notsurprised lots of geosocial news today w/ #s
xsw beginning - google fires a shot at foursquare with check in rewards {link}
4037                                  out of all my devices the ipad is the only one that
can hang an entire day at #sxsw... 37% remaining, the others died hours ago.
8828                                                      hmmû_a slew of iphone app upda
tes (inc. #4sq3) the past few days? can only mean one thing: it's #sxsw soon.
1550                                      @mention @mention @mention hope #apple visits
#art from the ipad, 8th &amp; congress since #sxsw gets own apple store {link}
5808                                                                 rt @mention google
circles is (not) a real thing and will (not) be launched today at #sxsw {link}
3318                                                                          i didn't
go to #sxsw because i'm still using an iphone 3g. #oldschool #novideo #veryslow
7859                                                             reid cites google's rou
te around as a good start to realtime/near realtime data {link} @mention #sxsw
5422                                         rt @mention apple has opened a pop-up store
in austin so that the nerds in town for #sxsw can get their new ipads. {link}
6414                                                      rt @mention one of my fav phot
os of #sxsw so far @mention &amp;  @mention #google #sxsw plixi.com/p/83881586
1878
video: ipad 2 line walk: austin texas. did you get one today? - #sxsw {link}
892                                                        hootsuite mobile for #sxsw ~
updates for iphone, blackberry &amp; android: shared by paulû_ {link} #shared
6490                 rt @mention rt @mention &quot;iava wants to be the google of nonprof
its.&quot; / yes, we do b/c our #vets deserve nothing less! #sxsw #letshookup
2933                                                                                  t
hanks @mention @mention for the fun party @mention for #sxsw last night: {link}
3259                                     google goggles + location could tell you, for e
xample, the history of a building you are looking at.  #augmentedreality #sxsw
8463                                             @mention they're flowing like water at
the pop up apple store in downtown austin near #sxsw. know anyone still there?
6725
rt @mention the #ipad 2 takes over #sxsw [video] - {link} #sxswi
4370                                           charity implications? rt @mention google
to launch major new social network called circles, possibly today {link} #sxsw
3715                                  good morning from #sxsw. who is standing in line f
or the ipad 2 today? i still have my ipad 1 for sale. 64gb wifi only $450. #fb
Name: text, dtype: object
```

In [104]:

```
# use a count vectorizer to get length and word counts for total vocabulary

cv = CountVectorizer()
cv_fit=cv.fit_transform(X)
word_list = cv.get_feature_names();
count_list = cv_fit.toarray().sum(axis=0)
word_df = pd.DataFrame()
word_df['words'] = word_list
word_df['counts'] = count_list
```

In [106]:

```
# total vocabulary: 9780 tokens

len(word_df)
```

Out[106]:

```
9780
```

In [108]:

```
# look at most common words
```

```
word_df.sort_values(by='counts', ascending=False).head(200)
```

Out[108]:

	words	counts
8334	sxsw	9628
5445	mention	7124
8574	the	4435
5067	link	4313
8714	to	3605
782	at	3105
7306	rt	2967
3751	google	2667
3409	for	2548
4570	ipad	2518
665	apple	2334
4384	in	1978
5978	of	1714
4606	is	1712
6860	quot	1696
582	and	1638
4583	iphone	1587
8140	store	1486
6019	on	1335
9066	up	1273
5812	new	1091
9641	you	1084
4620	it	1067
829	austin	973
576	an	873
9484	with	867
570	amp	836
5710	my	829
652	app	826
1705	circles	674
7872	social	667
4932	launch	653
8628	this	618
587	android	598
6529	pop	596
8716	today	584
1012	be	569
4759	just	559
3491	from	540
5890	not	536
8572	that	528

	words	counts
1417	by	526
6107	out	525
718	are	515
9646	your	484
5804	network	466
4572	ipad2	465
3991	have	439
9185	via	436
9447	will	418
5062	line	410
9338	we	405
321	about	399
3658	get	395
3464	free	390
5924	now	378
4316	if	362
1450	called	361
5393	me	357
6244	party	353
5578	mobile	352
7868	so	348
8359	sxswi	343
1466	can	339
8614	they	327
9394	what	326
514	all	322
1401	but	317
6022	one	317
6071	or	307
5265	major	304
5049	like	296
6925	re	296
3975	has	293
5868	no	289
4062	here	285
8678	time	275
8528	temporary	266
8605	there	258
6050	opening	257
1627	check	256
6567	possibly	244
9317	was	238
2291	day	234
6322	people	231
6104	our	230

	words	counts
2662	downtown	225
696	apps	225
7486	see	223
3813	great	222
5303	maps	220
5635	more	220
3715	go	218
3733	going	218
5377	mayer	218
4209	how	218
6044	open	215
760	as	212
6539	popup	211
5777	need	205
2608	don	199
2574	do	197
5318	marissa	193
4857	know	186
6027	only	179
1836	come	176
9749	ûï	175
3742	good	173
9455	win	172
8590	their	171
3307	first	168
3780	got	167
9092	us	165
8621	think	156
5820	news	156
1041	before	156
9416	who	155
2023	cool	151
5831	next	151
8489	tech	150
5692	music	150
9300	want	149
5179	love	146
6199	panel	145
1081	best	144
2395	design	143
7636	shop	142
8566	thanks	141
627	any	139
5104	ll	139
5119	location	137

	words	counts
3572	game	136
5267	make	136
885	awesome	136
1102	big	135
732	around	132
8563	than	132
7469	search	131
9099	use	130
7550	set	129
6121	over	129
9192	video	126
8737	too	126
9554	would	126
9619	year	126
2741	during	124
4914	last	124
8436	talk	123
7915	some	123
7655	show	122
9400	when	121
9106	users	121
9428	why	120
7226	right	119
631	anyone	118
3859	gt	117
9109	using	115
9150	ve	115
1121	bing	115
7392	says	114
4661	japan	114
7315	rumor	113
2653	download	113
6951	really	111
3888	guy	111
4629	its	110
3004	even	109
4935	launching	108
7548	session	108
8120	still	107
22	11	106
4536	into	104
4110	his	104
1934	congress	103
1844	coming	103
1407	buy	102

	words	counts
3538	future	102
1220	booth	101
8593	them	101
8943	twitter	100
1518	case	100
4070	hey	100
1831	com	99
246	6th	99
3128	facebook	98
9370	week	98
1143	blackberry	98
4635	itunes	98
6706	products	98
8480	team	97
2473	digital	97
4026	heard	95
9179	very	94
10	10	94
9333	way	94
9402	where	93
6374	phone	92
8505	technology	91
2651	down	91
2597	doing	90

Modeling

Model Evaluation

To evaluate all models, I split the data (not including the holdout set) into train and test sets, and found the recall and precision scores for the test set. The recall score shows what per cent of the true negative-sentiment tweets were captured by each model. The precision score shows what per cent of the tweets returned by the model are actually negative-sentiment. In order for Google to reduce their analysts' workload by half, precision must be at least 12%, since 6% of all tweets are negative-sentiment. Recall should be as high as possible.

I cross-validated the scores by testing each model using five different train/test splits. I then used the means of all five test set recall and precision scores as the final scores for that model.

Model 1: Baseline Model

The first model I tested used ScikitLearn's CountVectorizer to turn each tweet into a numerical vector by counting how many times each word appeared in the tweet. For this initial model, I included only single words and not n-grams in the vectorizer, and I did not set a maximum limit on the number of features. I then fed the vectorized tweets into ScikitLearn's Multinomial Naive Bayes classifier. This classifier supports binary as well as multi-class problems with discrete features, such as text classification.

The baseline model's **recall score was 0.12**, and its **precision score was 0.51**, as shown below. While this model would significantly speed up analysts' work as half of all tweets returned are truly negative-sentiment, it would not be useful to Google because about 90% of negative-sentiment tweets would be missed. The model is also overfit, suggesting it is relying too heavily on features of the training set.

In [25]:

```
# create count vectorizer for testing

countvec = CountVectorizer()
```

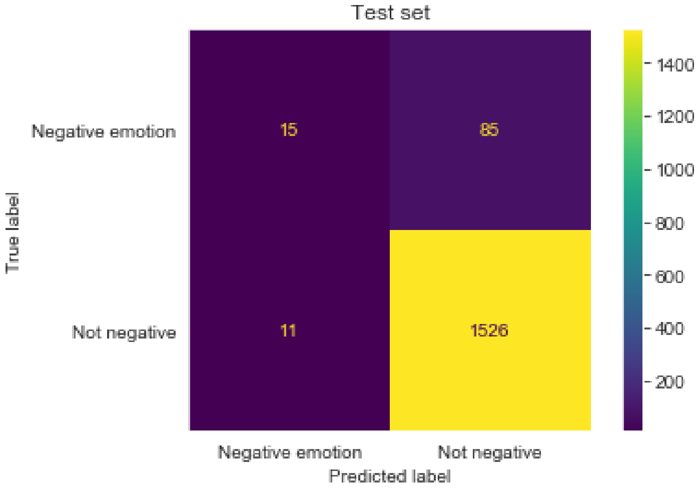
In [26]:

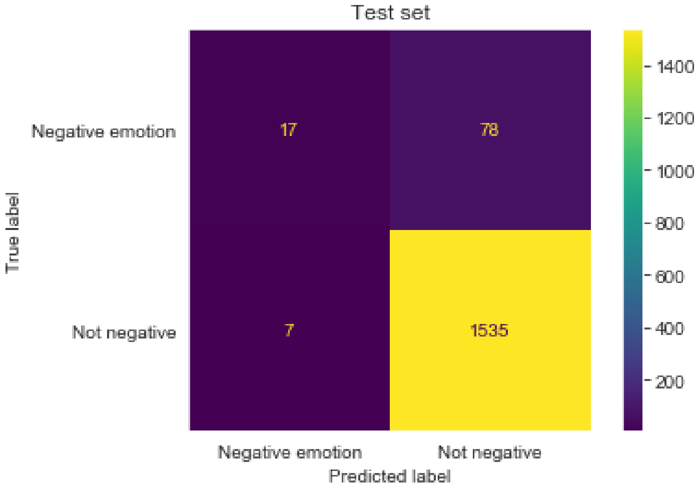
```
# create Multinomial Naive Bayes model for testing

multnb = MultinomialNB()
```

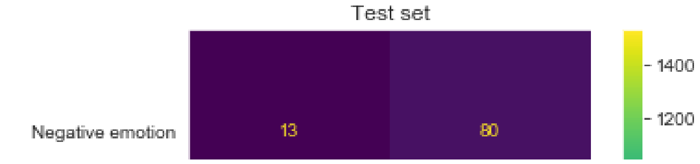
In [27]:

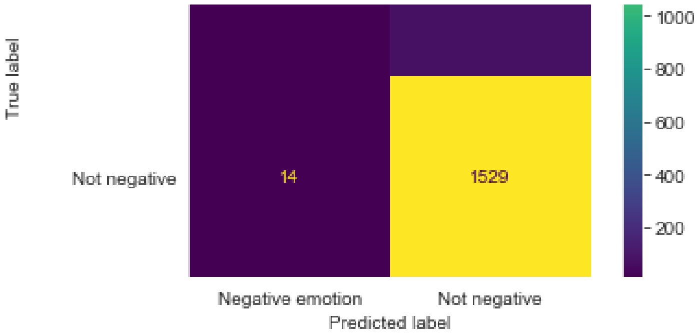
```
# test baseline model with basic countvectorizer and Multinomial Bayes model

ut.k_fold_validator(X_train, y_train, vectorizer=countvec, classifier=multnb, cv=5)
```

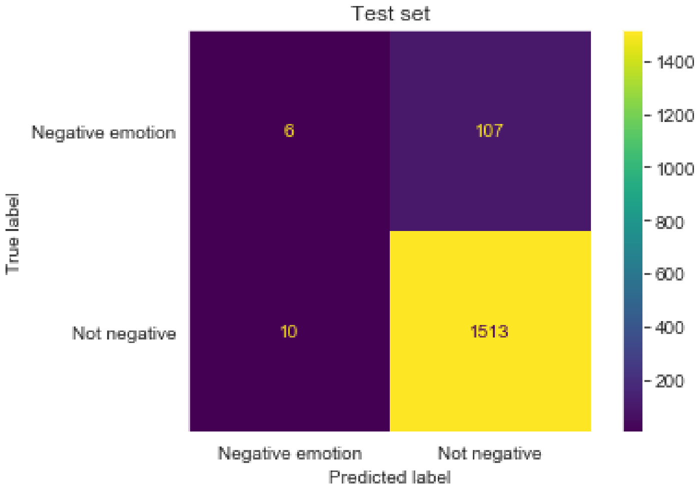
```
Vectorizer: CountVectorizer()
Classifier: MultinomialNB()
Cross-validation folds: 5


Train mean recall: 0.39 +/- 0.01
Train mean precision: 0.73 +/- 0.01
Train mean F1: 0.51 +/- 0.01


Test mean recall: 0.12 +/- 0.05
Test mean precision: 0.51 +/- 0.14
Test mean F1: 0.2 +/- 0.08
```

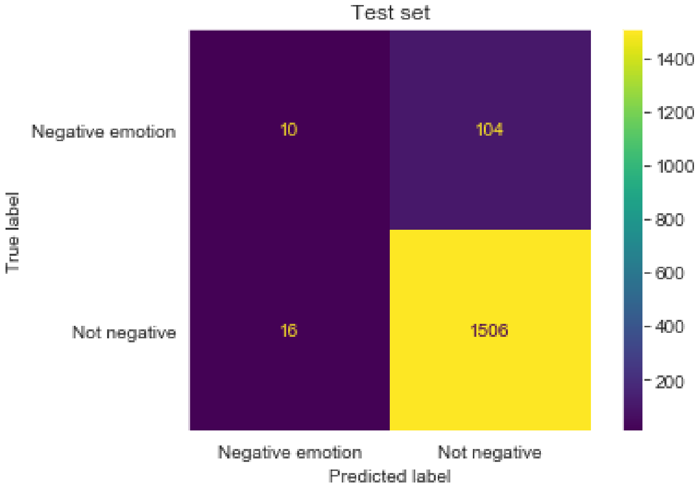












Model 2: TF-IDF Model

Next, I tested using a Term Frequency/Inverse Document Frequency vectorizer instead of the count vectorizer. This model produced even worse results, with a mean recall score of 0, meaning the model did not predict any tweets were negative-sentiment. A TF-IDF vectorizer downscales words that appear often in many documents to highlight words that truly typify one document. Tweets are so short that few words appear multiple times in any one tweet, and this is possibly why the model could not learn anything from the training data.

In [28]:

```
# create TF-IDF vectoriser for testing

tfidfvec = TfidfVectorizer()
```

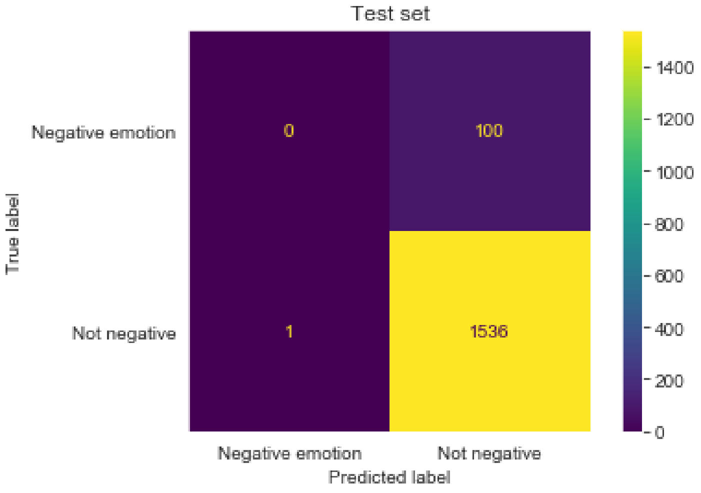
In [29]:

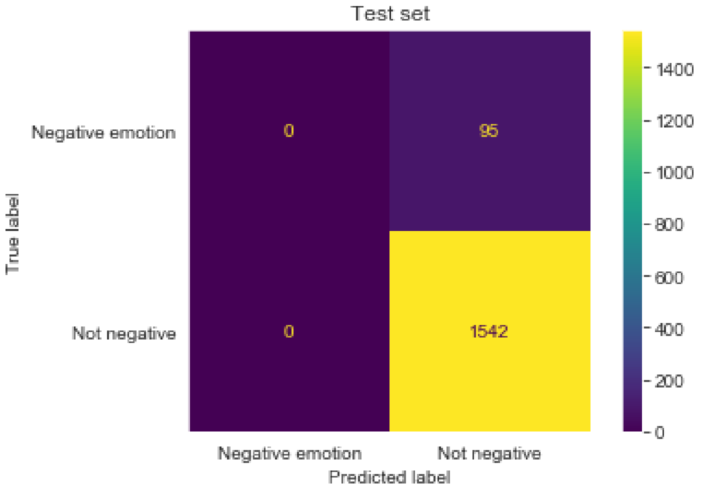
```
ut.k_fold_validator(predictor=X_train, target=y_train, vectorizer=tfidfvec, classifier=m
ultnb, cv=5)
```

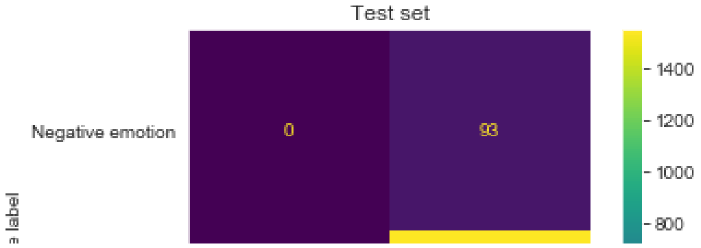
```
Vectorizer: TfidfVectorizer()
Classifier: MultinomialNB()
Cross-validation folds: 5
/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/sklearn/metrics/_classification
```

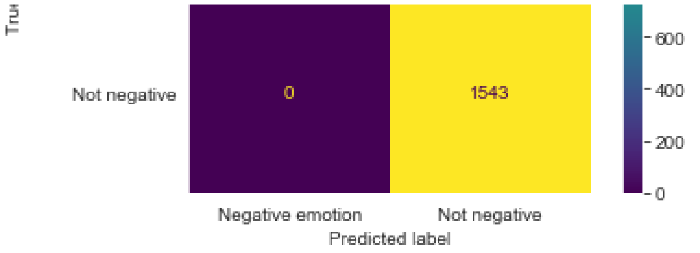
```
.py:1245: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 due to no
predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/sklearn/metrics/_classification
.py:1245: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 due to no
predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/sklearn/metrics/_classification
.py:1245: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 due to no
predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/sklearn/metrics/_classification
.py:1245: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 due to no
predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/opt/anaconda3/envs/learn-env/lib/python3.6/site-packages/sklearn/metrics/_classification
.py:1245: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 due to no
predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

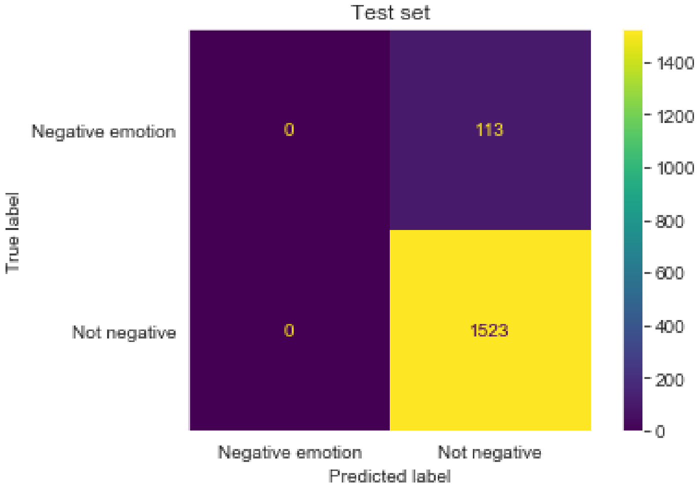
```
Train mean recall: 0.0 +/- 0.0
Train mean precision: 0.8 +/- 0.45
Train mean F1: 0.01 +/- 0.01


Test mean recall: 0.0 +/- 0.0
Test mean precision: 0.0 +/- 0.0
Test mean F1: 0.0 +/- 0.0
```

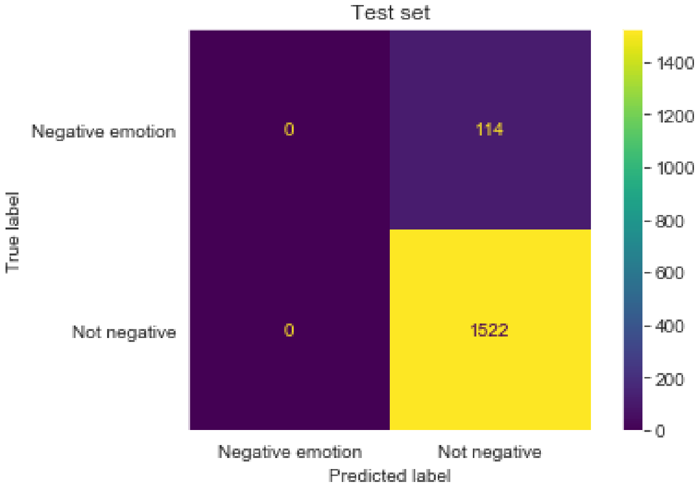












Model 3: Complement Naive Bayes Model

This model used a count vectorizer with a ScikitLearn's Complement Naive Bayes classifier. This classifier is designed to remedy class imbalance, which is a significant hurdle in this problem.

This model improved recall to 0.2, at the cost of precision which fell to 0.33. Since 0.34 is still above the 12% threshold set by Google, we will stick with this model to catch more negative-sentiment tweets.

In [30]:

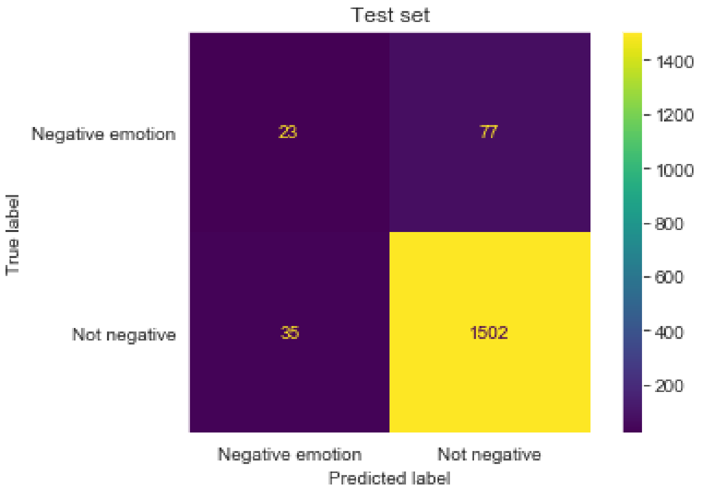
```
compnb = ComplementNB()
```

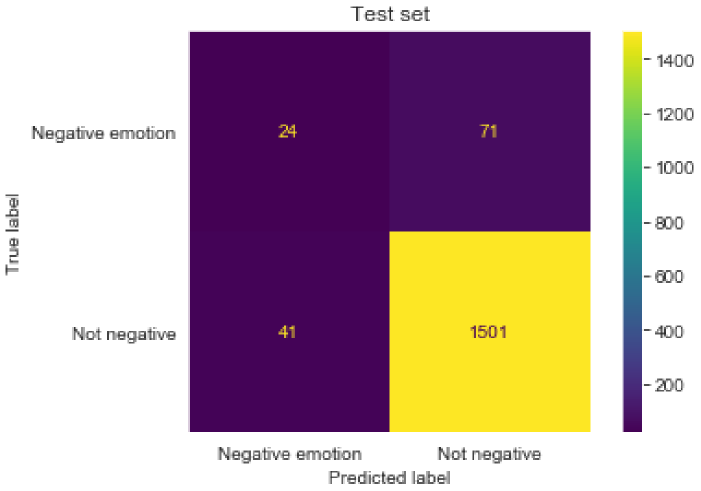
In [31]:

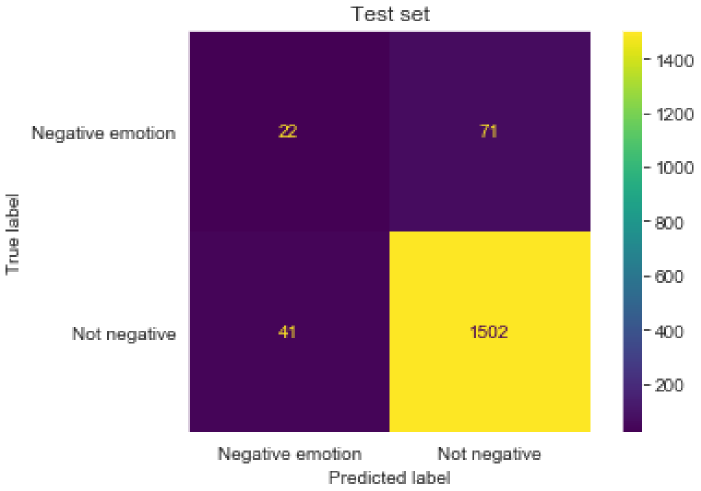
```
ut.k_fold_validator(predictor=X_train, target=y_train, vectorizer=countvec, classifier=compnb, cv=5)
```

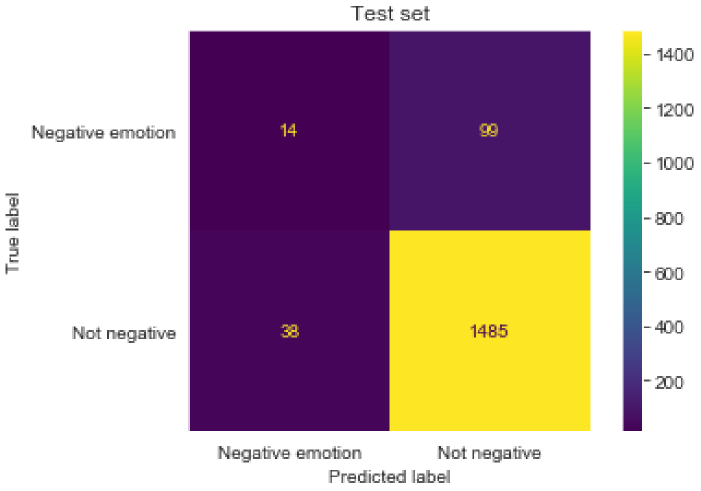
```
Vectorizer: CountVectorizer()
Classifier: ComplementNB()
Cross-validation folds: 5


Train mean recall: 0.58 +/- 0.02
Train mean precision: 0.51 +/- 0.01
Train mean F1: 0.54 +/- 0.01
```

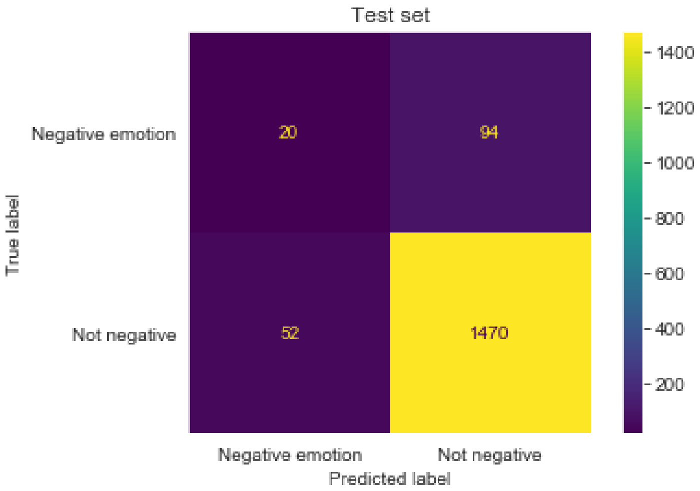
```
Test mean recall: 0.2 +/- 0.05
Test mean precision: 0.33 +/- 0.06
Test mean F1: 0.25 +/- 0.06
```











Model 4: Count Vectorizer with Stop Words

This model used a count vectorizer that contained a list of stop words - words to exclude from the vectorized data's features. This list included common English stopwords from Natural Language Toolkit, and punctuation. The model used a Complement Naive Bayes classifier.

Including these stopwords improved recall to 0.33, while precision fell to 0.28.

In [32]:

```
# create stopwords list for testing

stopwords_list = stopwords.words('english')
stopwords_list += string.punctuation
stopwords_list += ['sxsw','mention','rt']
```

In [33]:

```
stopwords_list
```

Out[33]:

```
['i',
 'me',
 'my',
 'myself',
 'we',
 'our',
 'ours',
 'ourselves',
 'you',
 "you're",
 "you've",
 "you'll",
 "you'd",
 'your',
 'yours',
 'yourself',
 'yourselves',
 'he',
 'him',
 'his',
 'himself',
 'she',
 "she's",
 'her',
 'hers',
 'herself',
 'it',
 "it's",
 'its',
 'itself',
 'they',
 'them',
 'their',
 'theirs',
 'themselves',
 'what',
 'which',
 'who',
 'whom',
 'this',
 'that',
 "that'll",
 'these',
 'those',
 'am',
 'is',
 'are',
 'was',
 'were',
 'be',
 'been',
 'being',
 'have',
 'has',
 'had',
 'having',
 'do',
 'does',
 'did',
 'doing',
 'a',
 'an',
 'the',
 'and',
 'but',
 'if',
 'or',
 'because',
 'as',
 'until',
 'while',
 'of',
 'at',
 'by',
 'for',
 'with',
 'about',
 'against',
 'between',
 'into',
 'through',
 'during',
 'before',
 'after',
 'above',
 'below',
 'to',
 'from',
 'up',
 'down',
 'in',
 'out',
 'on',
 'off',
 'over',
 'under',
 'again',
 'further',
 'then',
 'once',
 'here',
 'there',
 'when',
 'where',
 'why',
 'how',
 'all',
 'any',
 'both',
 'each',
 'few',
 'more',
 'most',
 'other',
 'some',
 'such',
 'no',
 'nor',
 'not',
 'only',
 'own',
 'same',
 'so',
 'than',
 'too',
 'very',
 's',
 't',
 'can',
 'will',
 'just',
 'don',
 "don't",
 'should',
 "should've",
 'now',
 'd',
 'll',
 'm',
 'o',
 're',
 've',
 'y',
 'ain',
 'aren',
 "aren't",
 'couldn',
 "couldn't",
 'didn',
 "didn't",
 'doesn',
 "doesn't",
 'hadn',
 "hadn't",
 'hasn',
 "hasn't",
 'haven',
 "haven't",
 'isn',
 "isn't",
 'ma',
 'mightn',
 "mightn't",
 'mustn',
 "mustn't",
 'needn',
 "needn't",
 'shan',
 "shan't",
 'shouldn',
 "shouldn't",
 'wasn',
 "wasn't",
 'weren',
 "weren't",
 'won',
 "won't",
 'wouldn',
 "wouldn't",
 '!',
 '"',
 '#',
 '$',
 '%',
 '&',
 "'",
 '(',
 ')',
 '*',
 '+',
 ',',
 '-',
 '.',
 '/',
 ':',
 ';',
 '<',
 '=',
 '>',
 '?',
 '@',
 '[',
 '\\',
 ']',
 '^',
 '_',
 '`',
 '{',
 '|',
 '}',
 '~',
 'sxsw',
 'mention',
 'rt']
```

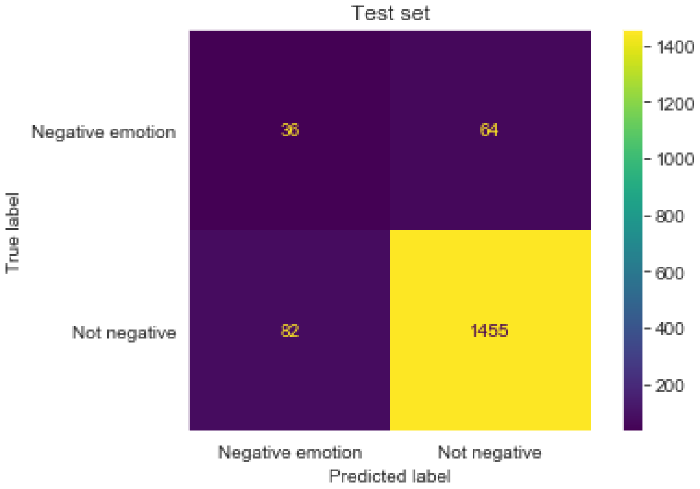
In [34]:

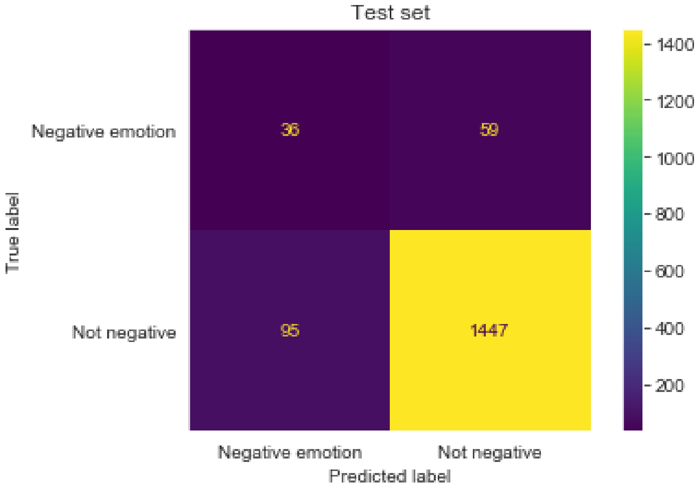
```
# create new count vectorizer that includes stopwords

cv_stop = CountVectorizer(stop_words=stopwords_list)
```

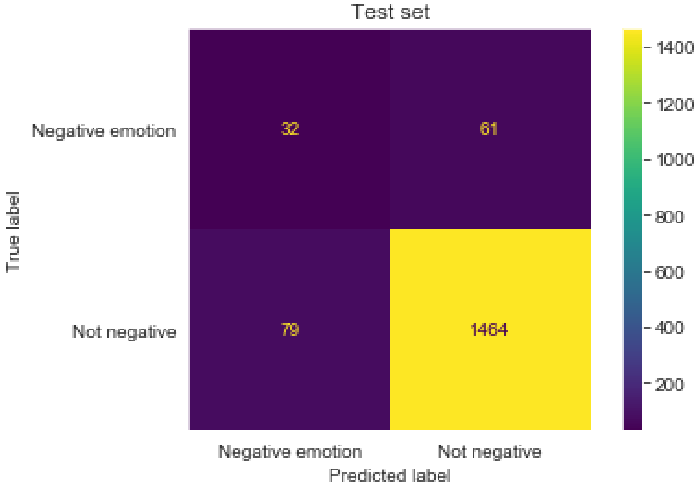
In [35]:

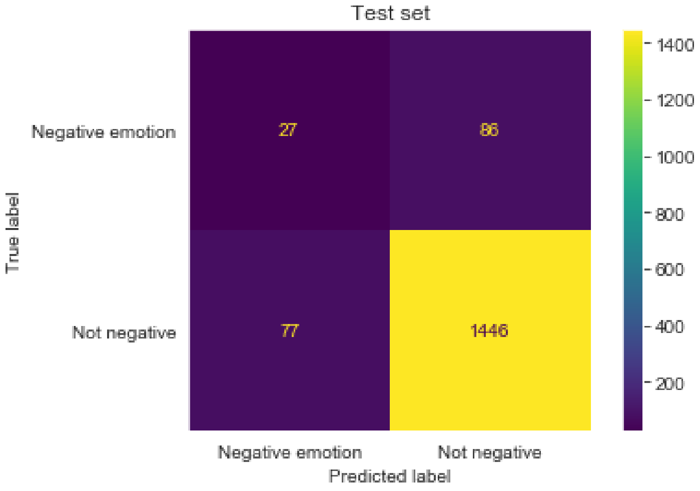
```
# test model 3

ut.k_fold_validator(predictor=X_train, target=y_train, vectorizer=cv_stop, classifier=compnb, cv=5)
```

```
Vectorizer: CountVectorizer(stop_words=['i', 'me', 'my', 'myself', 'we', 'our', 'ours',
                            'ourselves', 'you', "you're", "you've", "you'll",
                            "you'd", 'your', 'yours', 'yourself', 'yourselves',
                            'he', 'him', 'his', 'himself', 'she', "she's",
                            'her', 'hers', 'herself', 'it', "it's", 'its',
                            'itself', ...])
Classifier: ComplementNB()
Cross-validation folds: 5


Train mean recall: 0.72 +/- 0.01
Train mean precision: 0.42 +/- 0.01
Train mean F1: 0.53 +/- 0.01


Test mean recall: 0.33 +/- 0.06
Test mean precision: 0.28 +/- 0.02
Test mean F1: 0.3 +/- 0.03
```

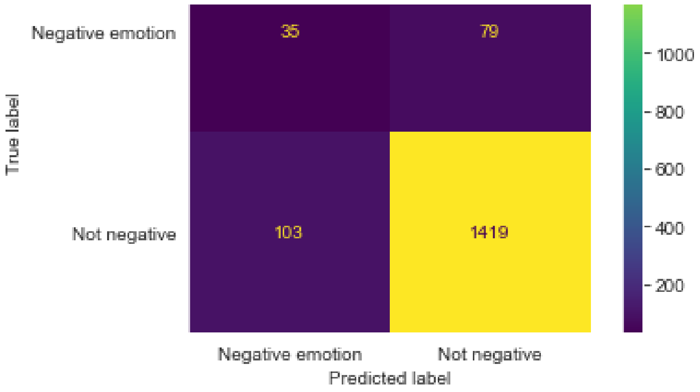

Test set


Test set


Test set


Test set

Test set



Model 5: Count Vectorizer with Short Stop Words List

Next, I tested a model with a very short stop words list, thinking that because tweets are brief, we may be getting good information from the small words and punctuation that had been excluded.

However, this stop words list did not improve model performance, as recall fell significantly to 0.23. Precision climbed only slightly to 0.31 but at the cost of missed negative-sentiment tweets. We will stick with the complete stop words list for now, but later will see that the shorter stop words list works better when using n-grams.

In [36]:

```
# create alternative stopwords list for testing
# since tweets are so short, we may be getting good info from punctuation and small words

stopwords_list_2 = ['sxsw', 'mention', 'rt']
```

In [37]:

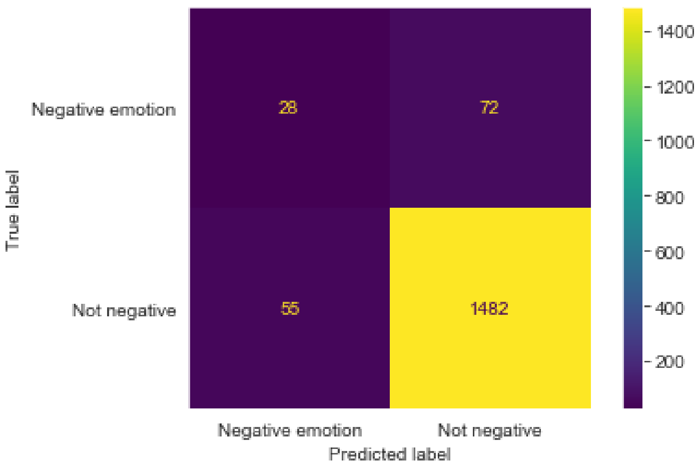
```
cv_stop_2 = CountVectorizer(stop_words=stopwords_list_2)
```

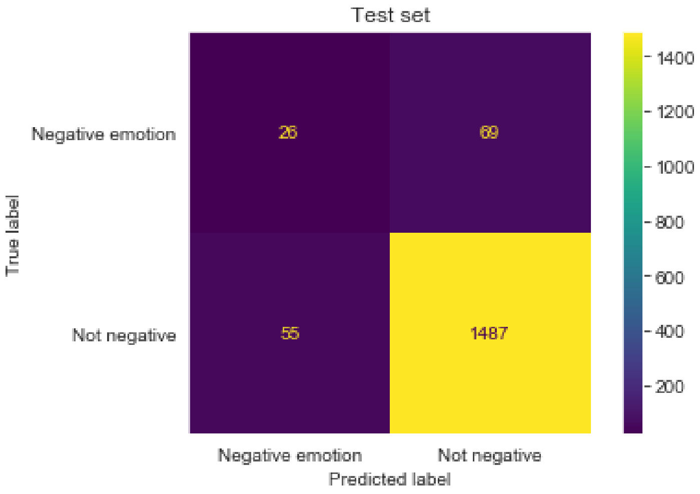
In [38]:

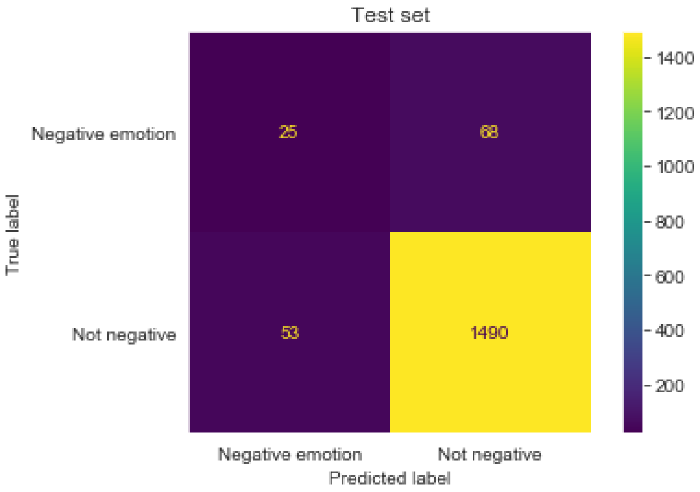
```
ut.k_fold_validator(predictor=X_train, target=y_train, vectorizer=cv_stop_2, classifier=
compnb, cv=5)
```

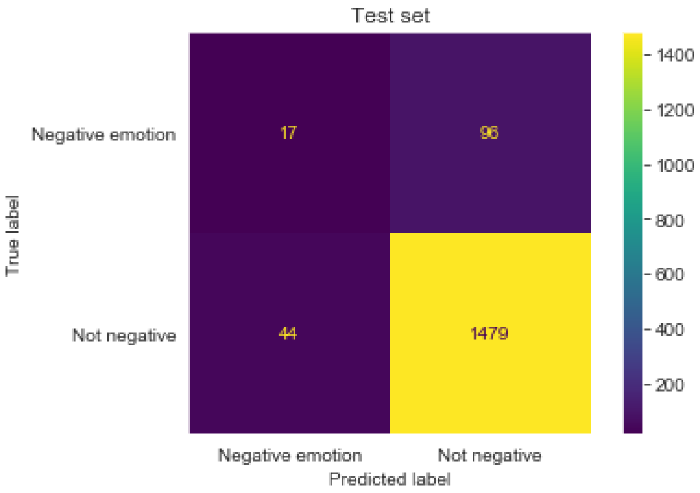
```
Vectorizer: CountVectorizer(stop_words=['sxsw', 'mention', 'rt'])
Classifier: ComplementNB()
Cross-validation folds: 5


Train mean recall: 0.61 +/- 0.02
Train mean precision: 0.48 +/- 0.01
Train mean F1: 0.54 +/- 0.01


Test mean recall: 0.23 +/- 0.06
Test mean precision: 0.31 +/- 0.03
Test mean F1: 0.26 +/- 0.05
```

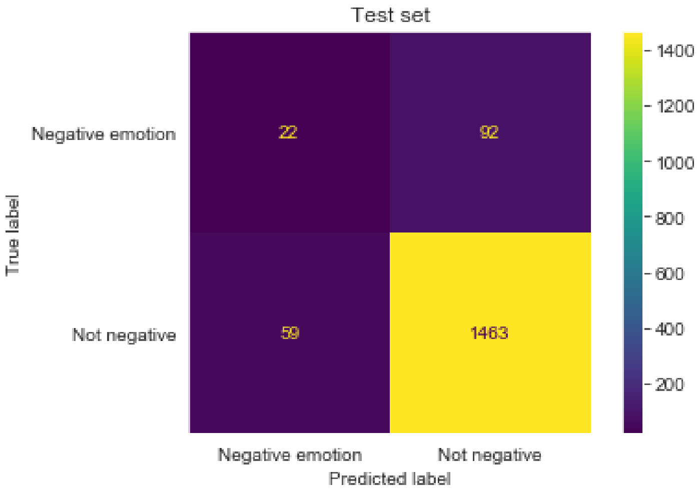
Test set











Model 6: Count Vectorizer with Stop Words and Max Features

This model set the maximum number of features to include in the vectorized data at 3,000.

Doing so significantly improved recall to 0.56 and reduced precision to 0.19, still above the threshold of 12%. Limiting the words to include to only the 3,000 most common words reduced overfitting in the model, and allowed it to classify more tweets as negative-sentiment.

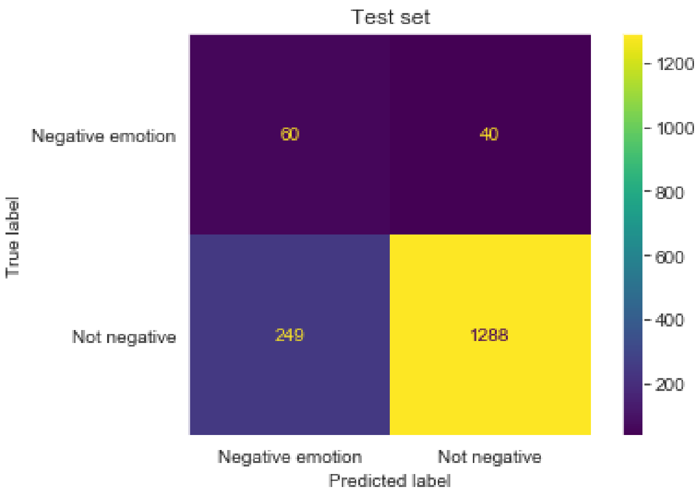
In [39]:

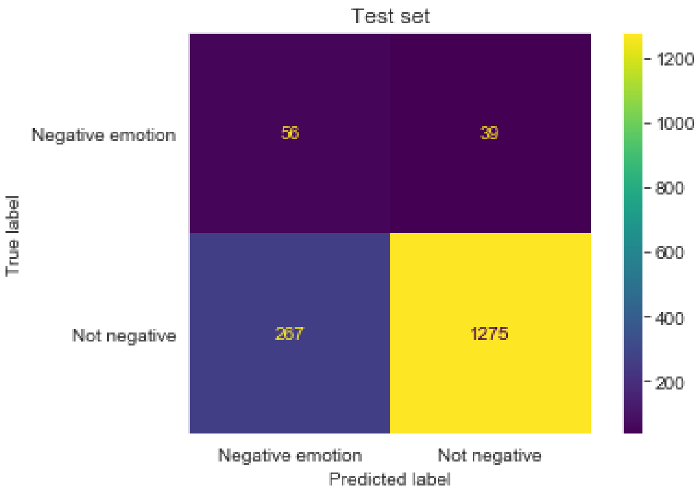
```
cv_stop_max = CountVectorizer(stop_words=stopwords_list, max_features=3000)
```

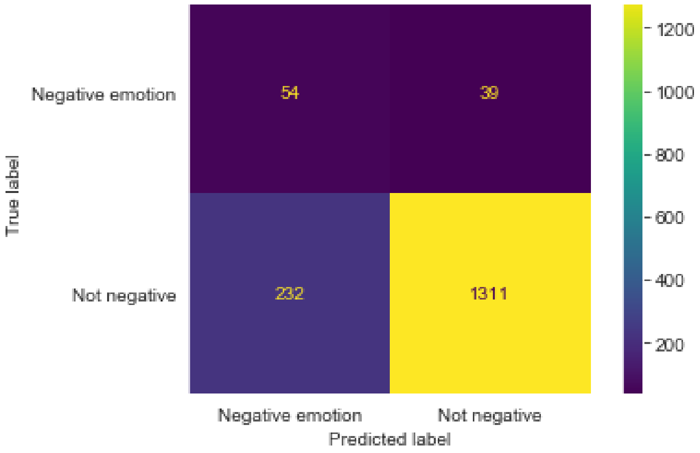
In [40]:

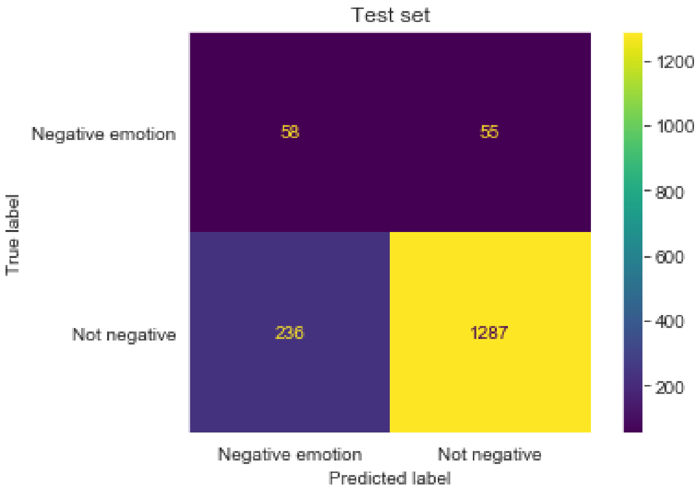
```
ut.k_fold_validator(predictor=X_train, target=y_train, vectorizer=cv_stop_max, classifie
r=compnb, cv=5)
```

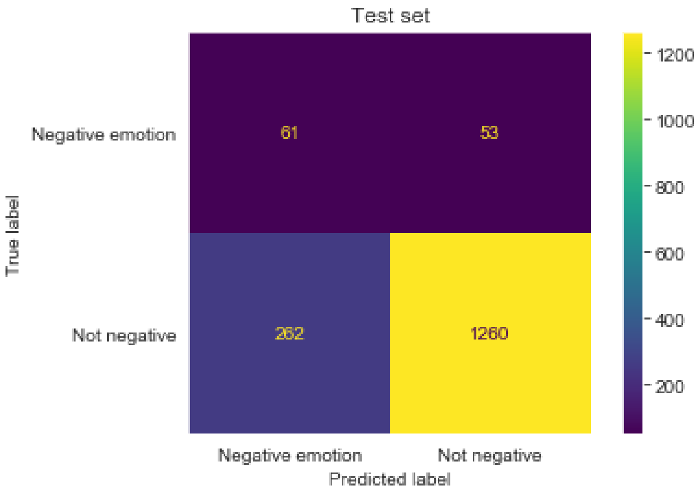
```
Vectorizer: CountVectorizer(max_features=3000,
                stop_words=['i', 'me', 'my', 'myself', 'we', 'our', 'ours',
                            'ourselves', 'you', "you're", "you've", "you'll",
                            "you'd", 'your', 'yours', 'yourself', 'yourselves',
                            'he', 'him', 'his', 'himself', 'she', "she's",
                            'her', 'hers', 'herself', 'it', "it's", 'its',
                            'itself', ...])
Classifier: ComplementNB()
Cross-validation folds: 5


Train mean recall: 0.82 +/- 0.01
Train mean precision: 0.25 +/- 0.01
Train mean F1: 0.39 +/- 0.01


Test mean recall: 0.56 +/- 0.04
Test mean precision: 0.19 +/- 0.01
Test mean F1: 0.28 +/- 0.01
```











Model 7: Count Vectorizer with Stop Words, Max Features, and N-grams

This model set a maximum n-gram length of 3, so single words as well as two- and three-word blocks would all be included as features. The model retained the stop words list and maximum feature limit used previously.

Recall jumped again to 0.65 while precision fell only slightly to 0.16, showing that these multi-word blocks are important features that the model can learn from.

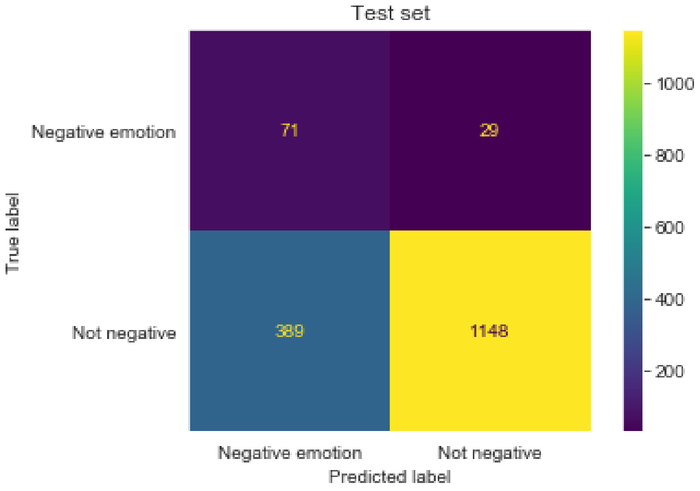
In [41]:

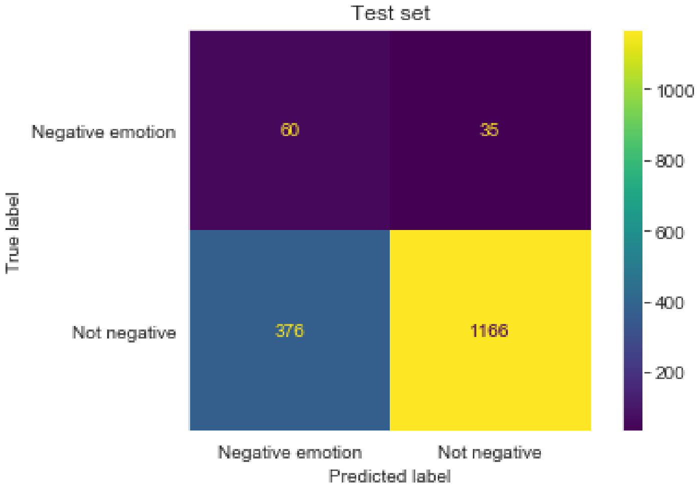
```
cv_stop_max_ngram = CountVectorizer(stop_words=stopwords_list, max_features=3000, ngram_range=(1, 3))
```

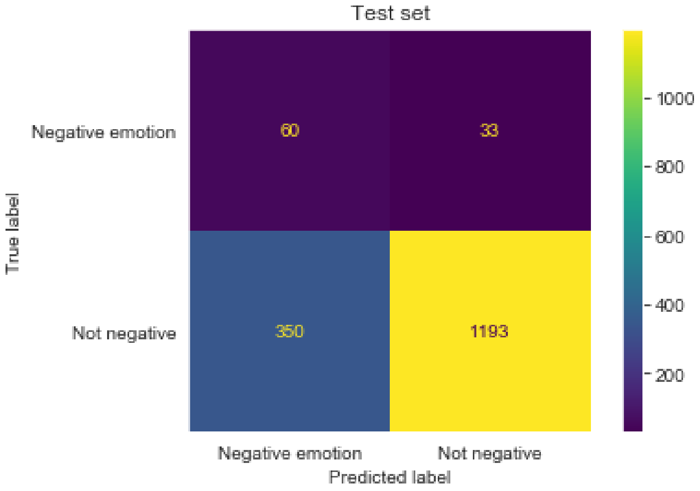
In [42]:

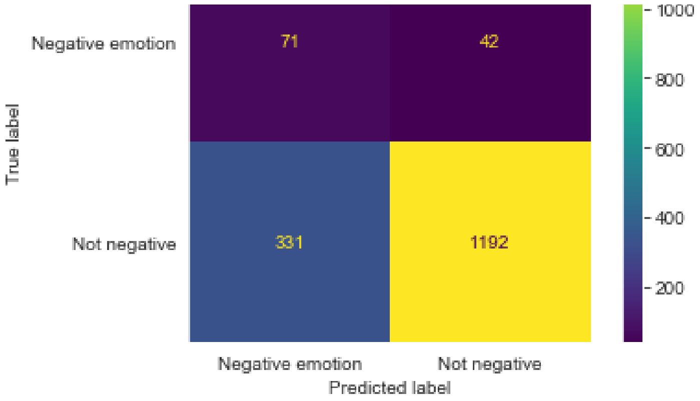
```
ut.k_fold_validator(predictor=X_train, target=y_train, vectorizer=cv_stop_max_ngram, classifier=compnb, cv=5)
```

```
Vectorizer: CountVectorizer(max_features=3000, ngram_range=(1, 3),
                stop_words=['i', 'me', 'my', 'myself', 'we', 'our', 'ours',
                            'ourselves', 'you', "you're", "you've", "you'll",
                            "you'd", 'your', 'yours', 'yourself', 'yourselves',
                            'he', 'him', 'his', 'himself', 'she', "she's",
                            'her', 'hers', 'herself', 'it', "it's", 'its',
                            'itself', ...])
Classifier: ComplementNB()
Cross-validation folds: 5


Train mean recall: 0.85 +/- 0.01
Train mean precision: 0.2 +/- 0.01
Train mean F1: 0.32 +/- 0.01


Test mean recall: 0.65 +/- 0.04
Test mean precision: 0.16 +/- 0.02
Test mean F1: 0.25 +/- 0.02
```

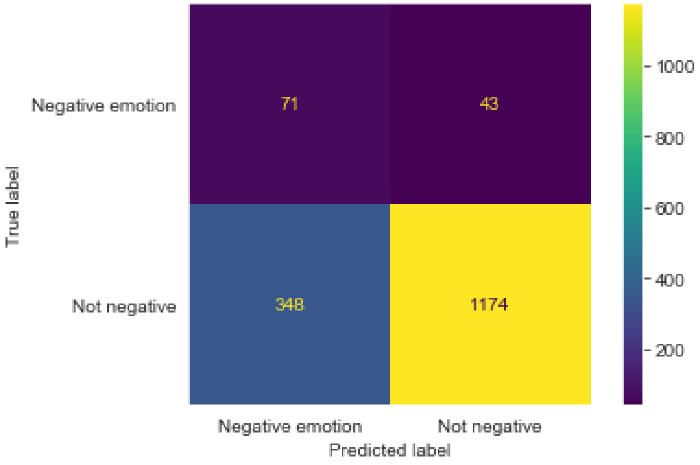








Test set



Model 8: Count Vectorizer with Stop Words, N-grams, and no Max Features

This model is identical to Model 7, but does not set a maximum feature limit.

I was curious to see the affect of n-grams without the feature limit. This model had a recall score of only 0.23 and a precision score of 0.45. N-grams only improve recall when a maximum feature limit is set. As with previous models lacking a maximum feature limit, this model was hugely overfit.

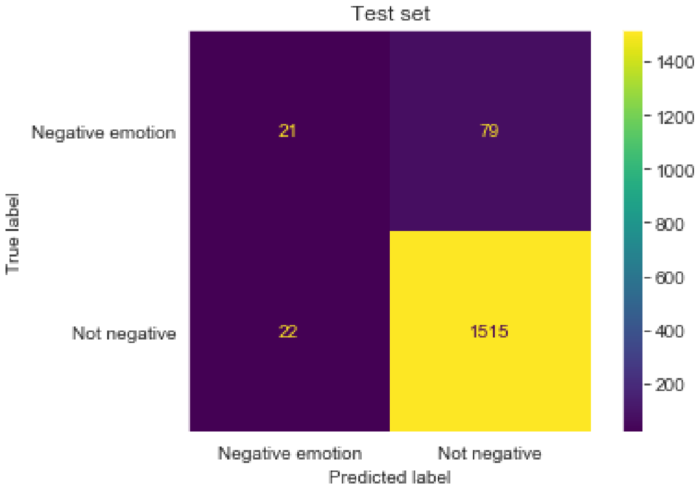
In [43]:

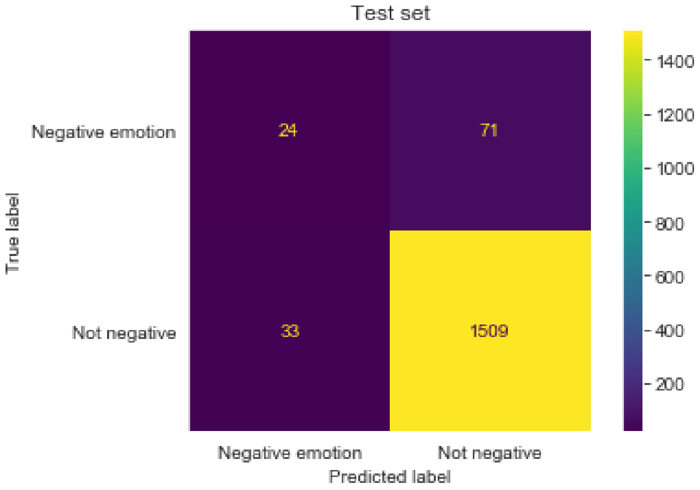
```
# create count vectorizer with stopwords list and ngrams but no max features limit

cv_stop_ngram = CountVectorizer(stop_words=stopwords_list, ngram_range=(1, 3))
```

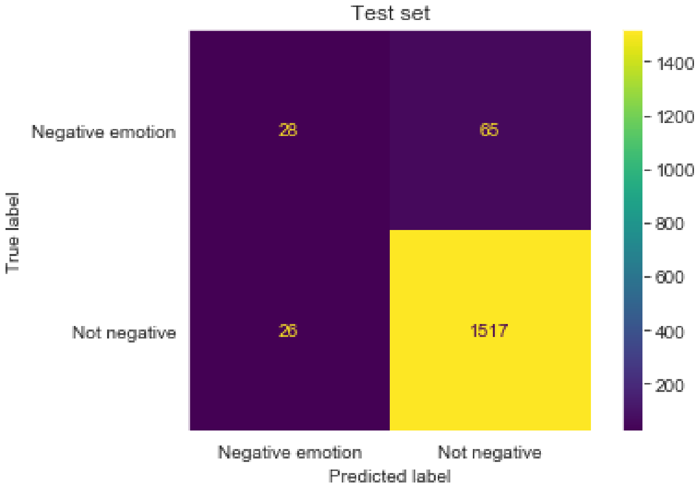
In [44]:

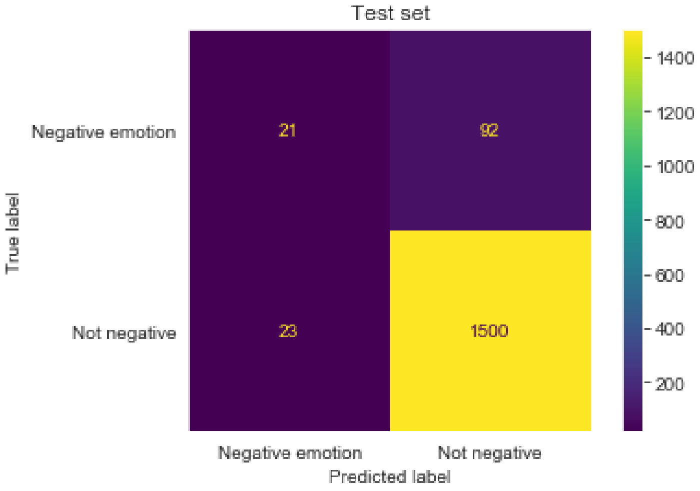
```
# ngrams don't improve recall without max

ut.k_fold_validator(predictor=X_train, target=y_train, vectorizer=cv_stop_ngram, classif
ier=compnb, cv=5)
```

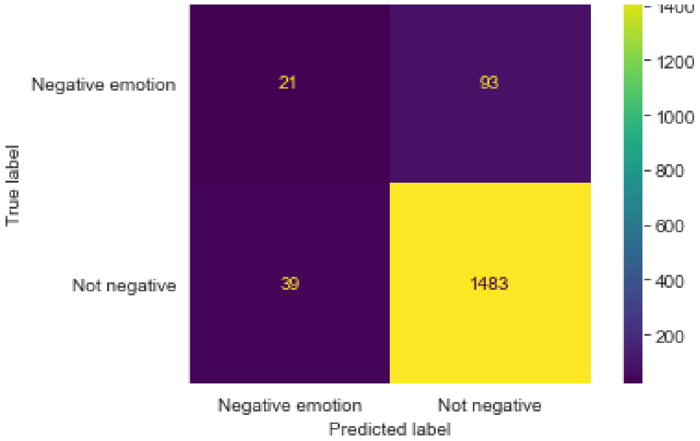
```
Vectorizer: CountVectorizer(ngram_range=(1, 3),
                stop_words=['i', 'me', 'my', 'myself', 'we', 'our', 'ours',
                            'ourselves', 'you', "you're", "you've", "you'll",
                            "you'd", 'your', 'yours', 'yourself', 'yourselves',
                            'he', 'him', 'his', 'himself', 'she', "she's",
                            'her', 'hers', 'herself', 'it', "it's", 'its',
                            'itself', ...])
Classifier: ComplementNB()
Cross-validation folds: 5


Train mean recall: 0.9 +/- 0.01
Train mean precision: 0.68 +/- 0.02
Train mean F1: 0.78 +/- 0.01


Test mean recall: 0.23 +/- 0.05
Test mean precision: 0.45 +/- 0.07
Test mean F1: 0.3 +/- 0.05
```











Model 9: Iterate on Maximum Features and N-gram Length

This model attempts to improve on Model 7 by testing maximum feature limits of 2000 and 4000, and n-gram maximum lengths of 2 and 4. After iterating through these combinations, a maximum feature limit of 2000 and an n-gram range between 1 and 3 performed best. A maximum feature limit of 1500 performed even better.

Recall jumped to 0.71 and precision fell slightly to 0.14 , still above the 12% threshold. This model is much less overfit, indicating that the smaller feature limit stops the model from learning too much from the training set.

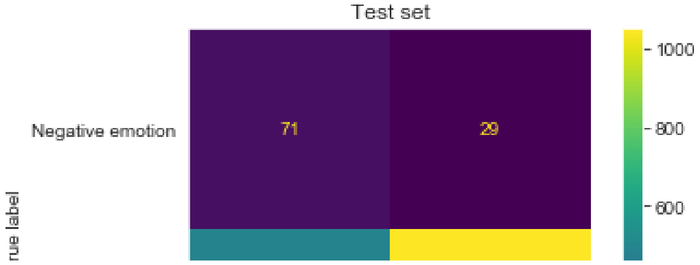
In [45]:

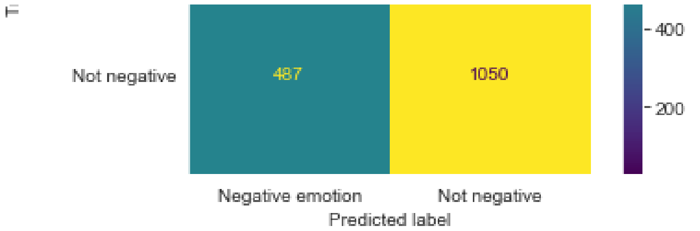
```
# best count vectorizer set maximum features at 1500, ngram range 1-3
# maximum features at 2k was a little better than 3k and 1k, and 1500 worked best
# maximum ngrams at 3 is better than 2 and 4
# minimum ngrams at 1 is better than 2

cv_stop_max_ngram_2 = CountVectorizer(stop_words=stopwords_list, max_features=1500, ngram
_range=(1, 3))
```

In [46]:

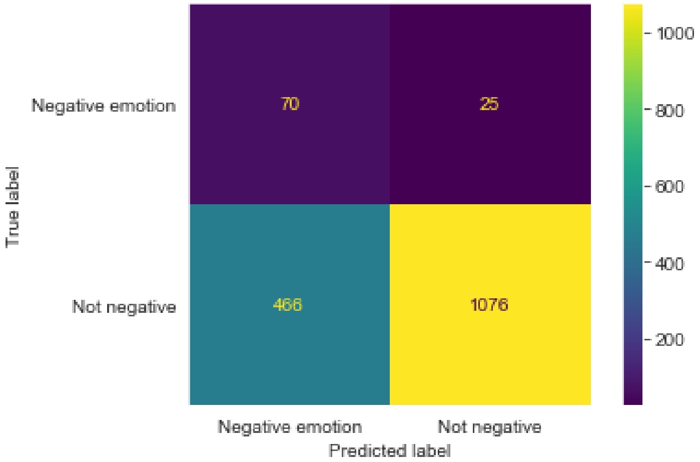
```
# test model with max features at 1500

ut.k_fold_validator(predictor=X_train, target=y_train, vectorizer=cv_stop_max_ngram_2, c
lassifier=compnb, cv=5)
```

```
Vectorizer: CountVectorizer(max_features=1500, ngram_range=(1, 3),
                stop_words=['i', 'me', 'my', 'myself', 'we', 'our', 'ours',
                            'ourselves', 'you', "you're", "you've", "you'll",
                            "you'd", 'your', 'yours', 'yourself', 'yourselves',
                            'he', 'him', 'his', 'himself', 'she', "she's",
                            'her', 'hers', 'herself', 'it', "it's", 'its',
                            'itself', ...])
Classifier: ComplementNB()
Cross-validation folds: 5


Train mean recall: 0.86 +/- 0.01
Train mean precision: 0.17 +/- 0.0
Train mean F1: 0.28 +/- 0.01


Test mean recall: 0.71 +/- 0.02
Test mean precision: 0.14 +/- 0.02
Test mean F1: 0.23 +/- 0.02
```

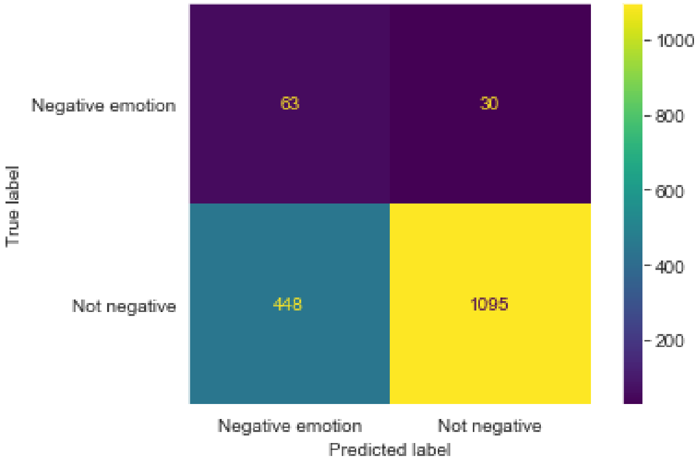




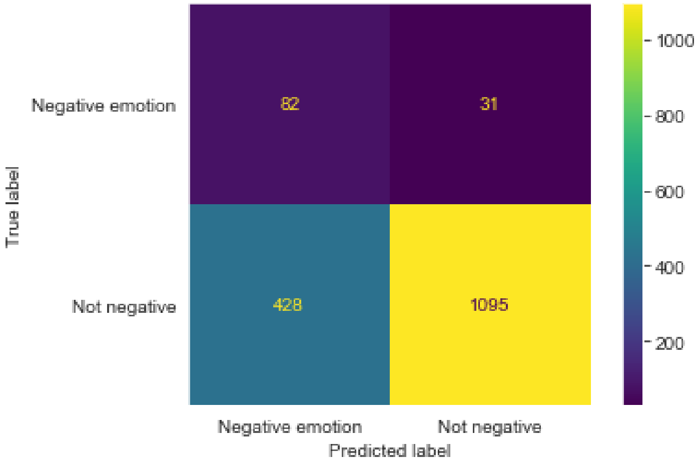
Test set



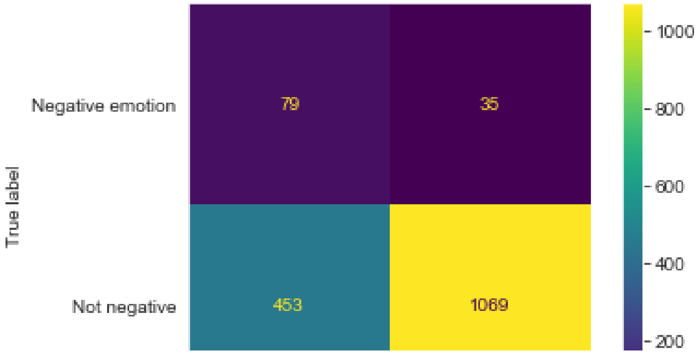
Test set



Test set



Test set



Model 10: Test Shorter Stop Words List with Current Best Model (# 9)

Since I only iterated on the stop words list before including n-grams, I was curious to see if the small words and punctuation I excluded might be useful now that two- and three-word n-grams are included. This model uses a list which only contains three very common words in these tweets: 'sxsw', 'rt', and 'mention'.

This change bumped up recall to 0.77 without sacrificing precision, which held steady at 0.14, indicating that when n-grams are included, common small words and punctuation are indeed useful. This model was the least overfit of any model tested, showing that it is learning only the information necessary to predict sentiment on new data.

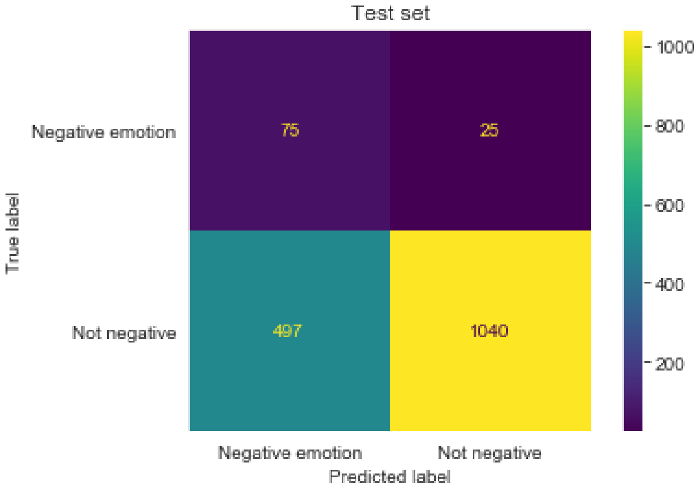
In [47]:

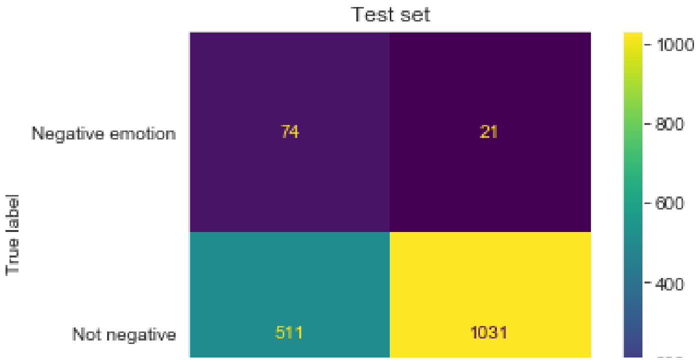
```
cv_stop_max_ngram_3 = CountVectorizer(stop_words=stopwords_list_2, max_features=1500, ngr
am_range=(1, 3))
```

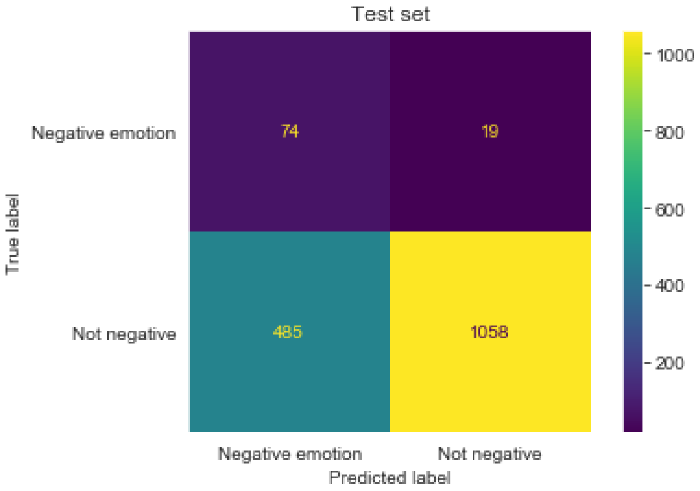
In [48]:

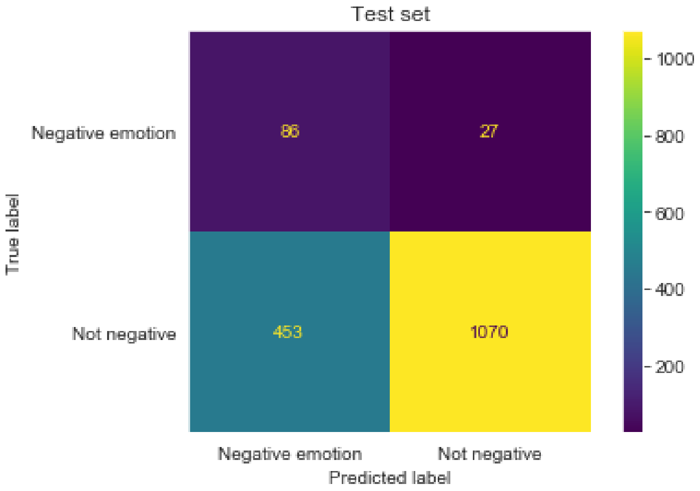
```
ut.k_fold_validator(predictor=X_train, target=y_train, vectorizer=cv_stop_max_ngram_3, c
lassifier=compnb, cv=5)
```

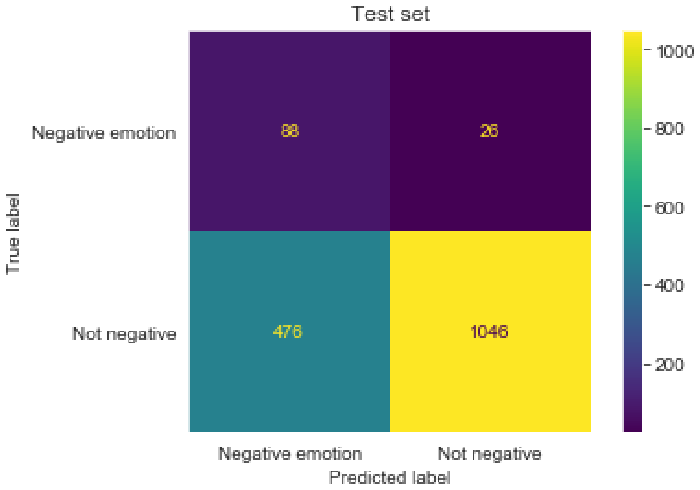
```
Vectorizer: CountVectorizer(max_features=1500, ngram_range=(1, 3),
                stop_words=['sxsw', 'mention', 'rt'])
Classifier: ComplementNB()
Cross-validation folds: 5


Train mean recall: 0.88 +/- 0.01
Train mean precision: 0.16 +/- 0.0
Train mean F1: 0.27 +/- 0.0


Test mean recall: 0.77 +/- 0.02
Test mean precision: 0.14 +/- 0.02
Test mean F1: 0.24 +/- 0.02
```











Model 11: Test No Stop Words List with Current Best Model (# 9)

This model is the same as Model 10, but includes no stop words list. **Recall was slightly worse at 0.75**, and **precision held steady at 0.14**, making Model 10 the best and final model.

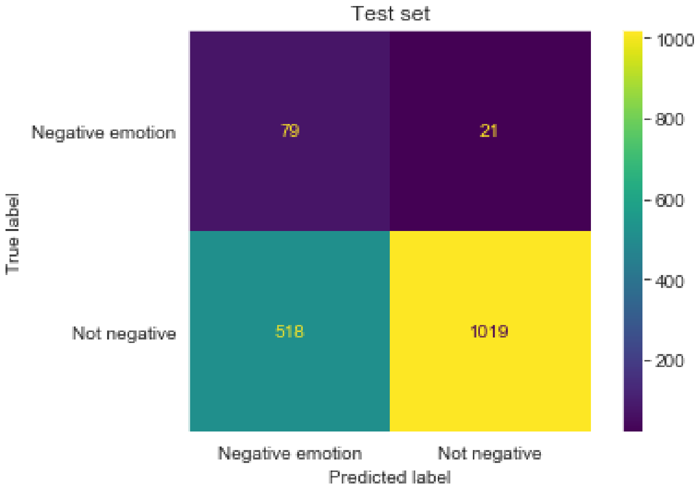
In [49]:

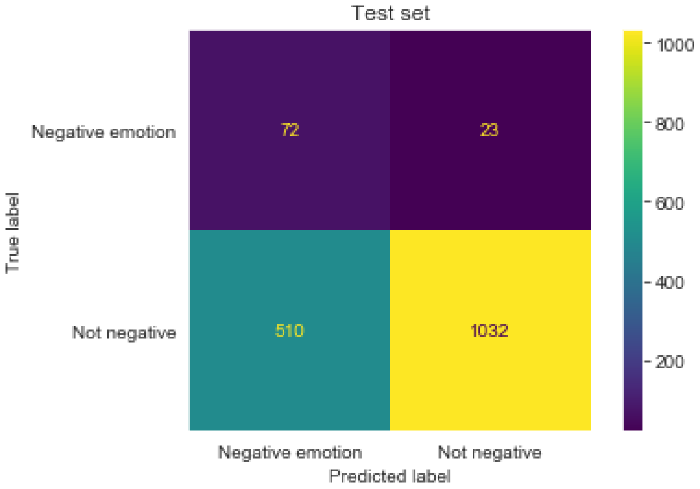
```
cv_max_ngram = CountVectorizer(max_features=1500, ngram_range=(1, 3))
```

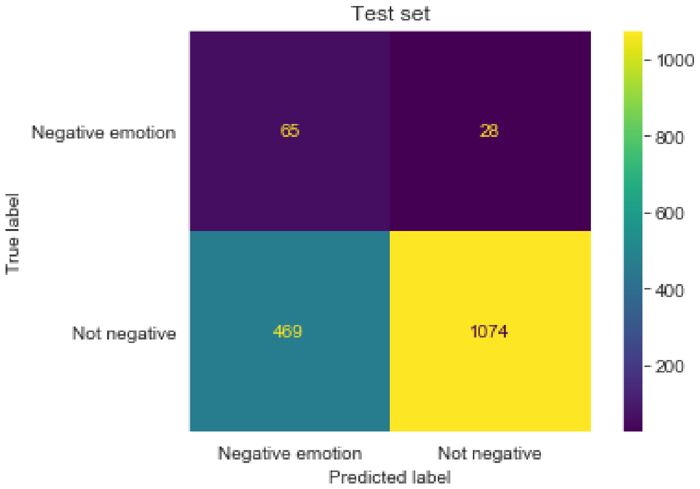
In [50]:

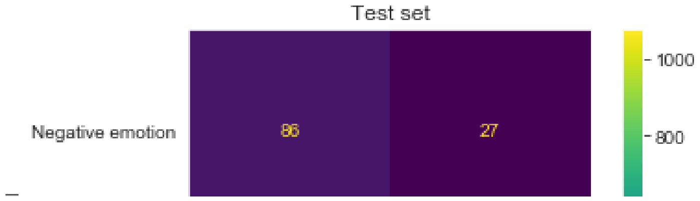
```
ut.k_fold_validator(predictor=X_train, target=y_train, vectorizer=cv_max_ngram, classifi
er=compnb, cv=5)
```

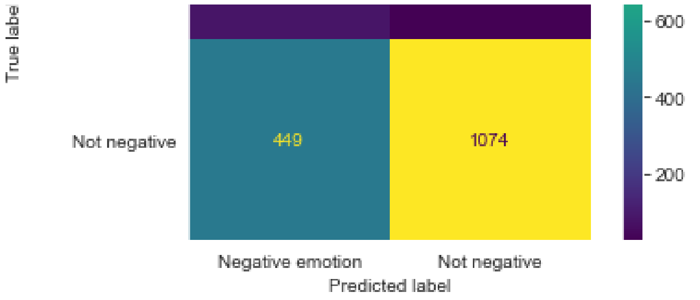
```
Vectorizer: CountVectorizer(max_features=1500, ngram_range=(1, 3))
Classifier: ComplementNB()
Cross-validation folds: 5


Train mean recall: 0.86 +/- 0.01
Train mean precision: 0.16 +/- 0.01
Train mean F1: 0.27 +/- 0.01


Test mean recall: 0.75 +/- 0.03
Test mean precision: 0.14 +/- 0.02
Test mean F1: 0.23 +/- 0.02
```

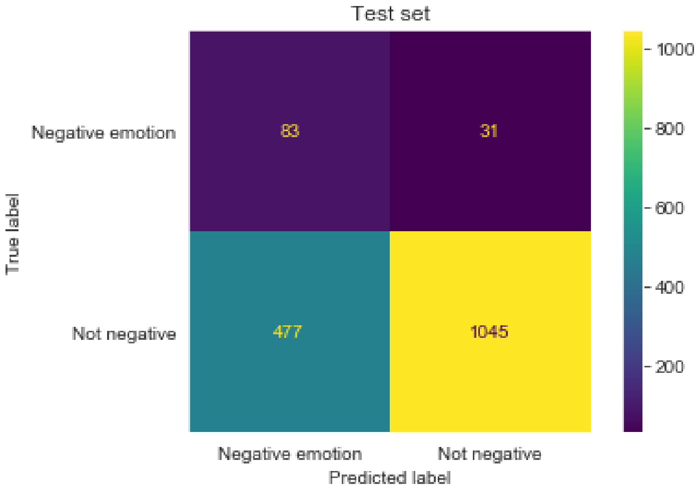











Test set

Final Model

The final natural language processing model includes the following features:

- A count vectorizer with:
  - Maximum feature limit of 1500
  - N-grams of between 1 and 3 words
  - A stop words list with only three words: 'sxsw', 'mention', and 'rt'
- A Complement Naive Bayesian classifier

Holdout Set Evaluation

As a final step, I ran the baseline model and the final model on the holdout set created at the beginning of the notebook to make sure the model could perform well on unseen data.

The model performed even better on the holdout set than it did during the testing process. **Recall was 0.87**, a full ten percentage points higher, and **precision was also higher at 0.15**. The model was not overfit, as the recall and precision for the training set were 0.88 and 0.16 respectively.

The model also performed better than the baseline model, which had recall and precision scores of 0.25 and 0.7 repsectively. A recall score of 0.25 would mean that too many negative tweets containing valuable information would go undetected.

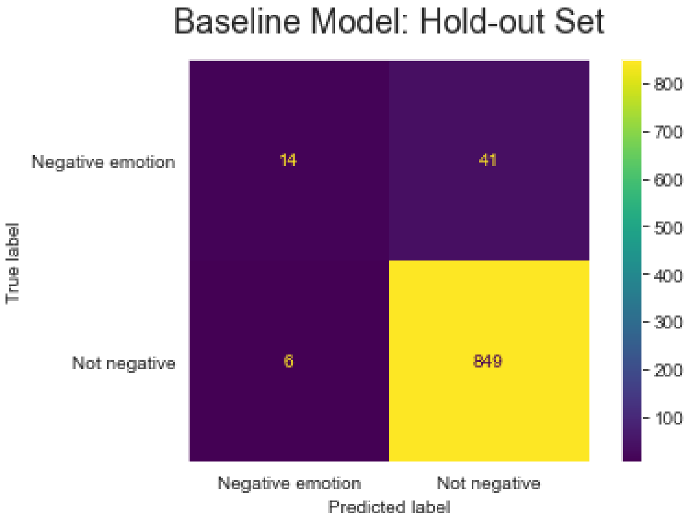
The improvement in performance on the holdout set vs the test sets used previously may be due to the fact that for this evaluation I trained the model using more data - the entire training set instead of just 75% of it as in the splits performed earlier. It is also possible that because the holdout set is small (just 10% of the original data), we got lucky and ended up with an easy data set.

In [51]:

```
# test baseline model on holdout set

vec = countvec
clf = multnb

X_vec_train = vec.fit_transform(X_train)
X_vec_holdout = vec.transform(X_holdout)
```

```
clf.fit(X_vec_train, y_train)

y_pred_train = clf.predict(X_vec_train)
y_pred_holdout = clf.predict(X_vec_holdout)

print('Train recall score:', round(recall_score(y_train, y_pred_train, pos_label='Negati
ve emotion'), 2))
print('Train precision score:', round(precision_score(y_train, y_pred_train, pos_label='
Negative emotion'), 2))
print('Train F1 score:', round(f1_score(y_train, y_pred_train, pos_label='Negative emotio
n'), 2))
print('\n')
print('Holdout recall score:', round(recall_score(y_holdout, y_pred_holdout, pos_label='N
egative emotion'), 2))
print('Holdout precision score:', round(precision_score(y_holdout, y_pred_holdout, pos_la
bel='Negative emotion'), 2))
print('Holdout F1 score:', round(f1_score(y_holdout, y_pred_holdout, pos_label='Negative
emotion'), 2))

plot_confusion_matrix(clf, X_vec_holdout, y_holdout)
plt.title('Baseline Model: Hold-out Set', fontsize=18, pad=15)
plt.savefig('images/baseline-model-holdout')
```

```
Train recall score: 0.42
Train precision score: 0.69
Train F1 score: 0.53


Holdout recall score: 0.25
Holdout precision score: 0.7
Holdout F1 score: 0.37
```

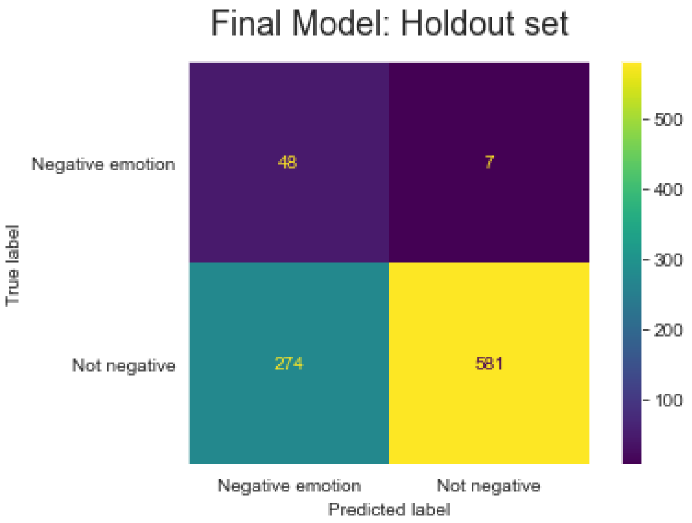


In [52]:

```
# test final model on holdout set

vec = cv_stop_max_ngram_3
clf = compnb

X_vec_train = vec.fit_transform(X_train)
X_vec_holdout = vec.transform(X_holdout)

clf.fit(X_vec_train, y_train)

y_pred_train = clf.predict(X_vec_train)
y_pred_holdout = clf.predict(X_vec_holdout)

print('Train recall score:', round(recall_score(y_train, y_pred_train, pos_label='Negati
ve emotion'), 2))
print('Train precision score:', round(precision_score(y_train, y_pred_train, pos_label='
Negative emotion'), 2))
print('Train F1 score:', round(f1_score(y_train, y_pred_train, pos_label='Negative emotio
n'), 2))
```

```
print('\n')
print('Holdout recall score:', round(recall_score(y_holdout, y_pred_holdout, pos_label='Negative emotion'), 2))
print('Holdout precision score:', round(precision_score(y_holdout, y_pred_holdout, pos_label='Negative emotion'), 2))
print('Holdout F1 score:', round(f1_score(y_holdout, y_pred_holdout, pos_label='Negative emotion'), 2))

plot_confusion_matrix(clf, X_vec_holdout, y_holdout)
plt.title('Final Model: Holdout set', fontsize=18, pad=15)
plt.savefig('images/final-model-holdout')
```

```
Train recall score: 0.88
Train precision score: 0.16
Train F1 score: 0.27


Holdout recall score: 0.87
Holdout precision score: 0.15
Holdout F1 score: 0.25
```



Conclusions

Identifying negative-sentiment tweets is a challenging problem since they comprise just 6% of all tweets in the dataset. The final model provides value to Google by enabling analysts to work over twice as fast, while still catching 77% of available negative-sentiment tweets.

The model's precision score was 0.14, meaning that 14 out of every hundred tweets returned by the model are truly negative-sentiment. Without the model, analysts would find only 6 negative tweets in every one hundred. The model's 77% recall rate means that most negative-sentiment tweets would be captured by the model. Since each tweet contains valuable information that can help Google understand customer frustrations, the company would like to capture as many of these tweets as possible.

The final model improved on the baseline model, which identified only 12% of all negative-sentiment tweets, though about half the tweets it returned truly were negative-sentiment.

Future Work

Natural language processing is a complex area of machine learning that has many different tools available for data scientists. Testing additional vectorizers and models may improve on these results.

While a count vectorizer worked well for this problem, the weakness of a count vectorizer is that it does not account for the meanings of words. Moreover, words that are not in the training set cannot help drive predictions even if they are in the test set. Pre-trained vectorizers such as Google's Word2Vec, Stanford's GloVe, and SpaCy may produce better results.

There are several other model types which would be interesting to test with this problem. Decision trees and support vector machines are robust classification algorithms and may work well for this problem. It would also be interesting to test a neural net.