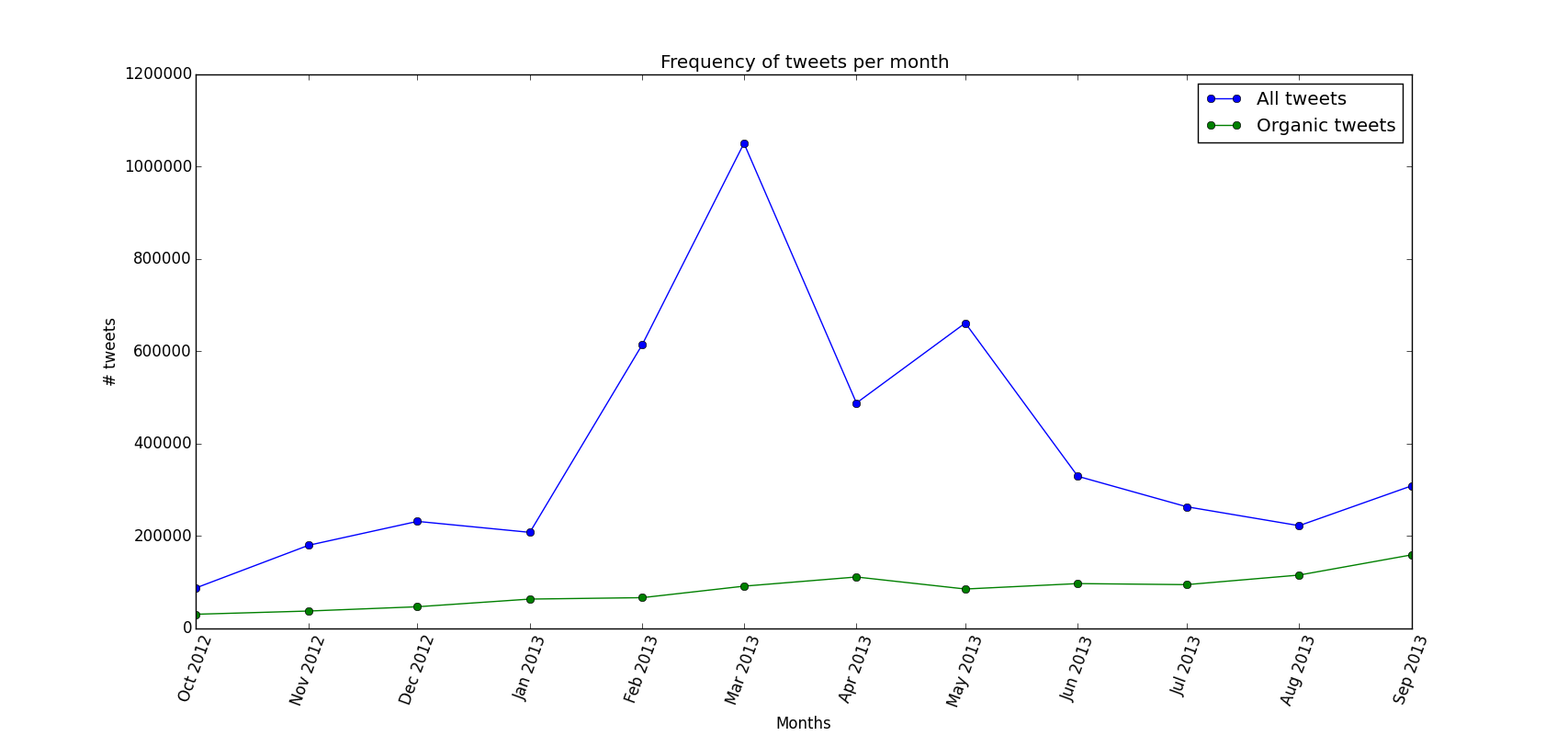
Data

The data is composed for 4,639,885 tweets which used hashtags about electronic cigars, and it includes personal tweets (we call them organic) and tweets from companies’ profiles. In order to work just with organic tweets we have done a preprocessing that splits the data, and we got just the organic tweets that totalize 992633 tweets. The plot below shows the number of tweets (y axis) by month (x axis). The green line refers to organic tweets and blue line to all tweets

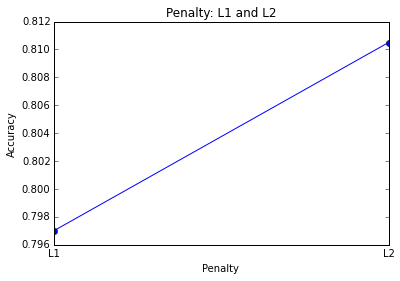
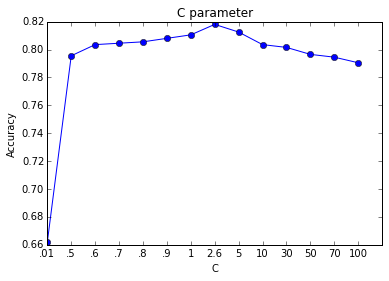
After this we have randomly selected 2000 from the set of organic tweets and then we manually labeled these tweets as positive, negative and neutral.

We have joined the negative and neutral because we are more interested in positive than negative and also because the accuracy got higher with just considering positive and negative. So, at this point we have 676 classified as positive and 1324 as negative.

Classification task

Training

We have created a Logistic Regression (LG) classifier, which is responsible to predict the labels of all the organic tweets. The labeled tweets were used as our training set. We have used a grid search to check the best values for ‘Penalty’ and ‘C’, which are parameters of LG, in order to have a better accuracy for the classifier. Running the grid search we have found out that L2 and C=2.6 are the best configuration for LG. The plots below show that the accuracy is higher when those parameters are chosen.

Our best classifier configuration has the accuracy of 81.8%. The confusion matrix of the model is shown in the Figure 1 and right below of it precision, recall and F-score values are described for both classes (negative and positive). As we can notice the percentage retrieved tweets that are relevant is 90% what is seen as a good value for our problem.

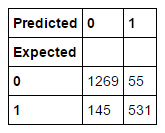


Figure 1 - Confusion Matrix of Logistic Regression Classifier

Precision:[0.89745403, 0.90614334])

Recall:[0.95845921, 0.78550296])

F-score:[0.92695398, 0.84152139])

Testing –

Our test set is defined as all organic tweets (900k tweets). After running the classifier on them we had 27% classified as positive and 73% as negative. A new LG model was built with the classified tweets in order to double check the accuracy of our model and them new 200 tweets were randomly selected and labeled to be our testing set and we had 79% of accuracy for those. We conclude that our model is good for predicting the sentiment of the tweets.

After we have decided to do the analysis with this model we started to check the top and bottom words of our class. They are shown below:

**Top coefficients:**

[(u'THIS\_IS\_A\_URL', -55.800122860792492), (u'cigarettes', -26.807477330349069), (u'retail', -21.400333836770933), (u'stores', -20.681826645875173), (u'he', -19.342041543260226), (u'everyone', -19.070127486024173), (u'smokes', -18.084708661458365), (u'you', -17.434688780794623), (u'dallas', -17.097809505370368), (u'as', -17.07282274807854), (u'kid', -16.684916367926963), (u'go', -16.650099465536549), (u'if', -16.2490745280782), (u'is', -16.089329425068758), (u'ego', -15.757535240901543), (u'people', -15.693700004959279), (u'being', -15.565127121243144), (u'bro', -15.430834066412707), (u'any', -15.124662561973045), (u'many', -14.983983850516712), (u'one', -14.911291961091866), (u'even', -14.798283429009391), (u'have', -14.490123743413086), (u'ecigarettes', -14.395299898135512), (u'via', -14.209716487838863), (u'hate', -14.181269156485163), (u'she', -14.14817941860678), (u'caught', -13.722834662189943), (u'video', -13.663875051096152), (u'by', -13.605663769072526)]

**Bottom coefficients:**

[(u'since', 11.565339764605838), (u'and', 11.566483625681062), (u'euecigban', 11.84392352692768), (u'flavoured', 11.898545774517881), (u're', 12.083784864364034), (u'anyone', 12.09821734613946), (u'black', 12.343631368992128), (u'alternative', 12.43366116742714), (u'with', 12.562267335195829), (u'drag', 12.573129036833173), (u'quit', 12.595324510412709), (u'this', 13.251940046695436), (u'vapes', 13.510995012969385), (u'ecig', 13.544450385701102), (u'gotta', 14.196731763899063), (u'cig', 14.57517127138235), (u'good', 15.052673261593087), (u'try', 15.430016647575419), (u'tastes', 15.92362083947822), (u'got', 16.13278035035885), (u'me', 16.254375159020366), (u'taste', 16.408611436225481), (u'smokin', 16.677182881059156), (u'we', 17.277320176715904), (u'buy', 17.798299700492532), (u'green', 20.568172009523487), (u'vape', 28.606092110683399), (u'vaping', 31.956322540259507), (u'i', 60.075229642053195), (u'my', 74.542920034164979)]

**Analysis of the sentiment during a year (Oct/2012-Sep/2013)**

We have analyzed the amount of tweets by month according to the sentiment (negative or positive) toward e-cigars. In the Figure 2 we have considered all organic tweets, in the Figure 3 it is considered just one tweet by user and in the Figure 4 it is not considered retweets.

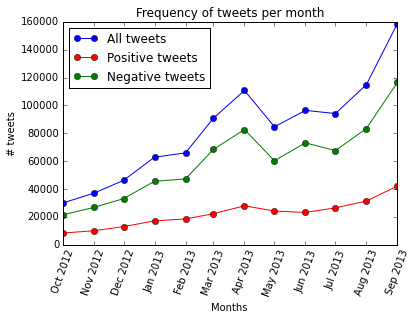


Figure 2 - all organic

Figure 2 shows spikes in April, June and September in all categories but Figure 3, which represents one tweet per user, has moved the first left spike from April to March, what tell us that people have posted more tweets in April. Figure 3 and Figure 4 follow the same shape they differ just in the number of tweets.

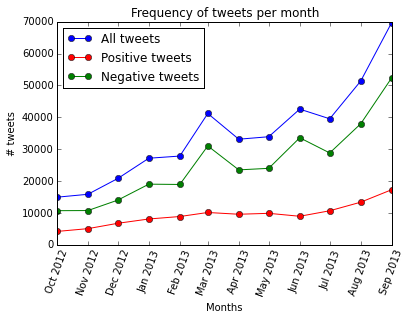


Figure 3 - one tweet per user

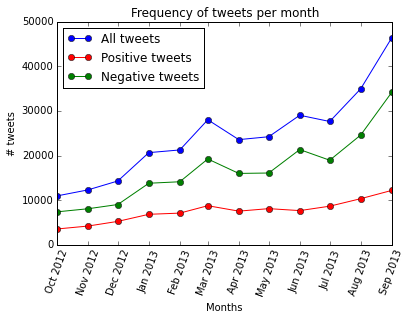
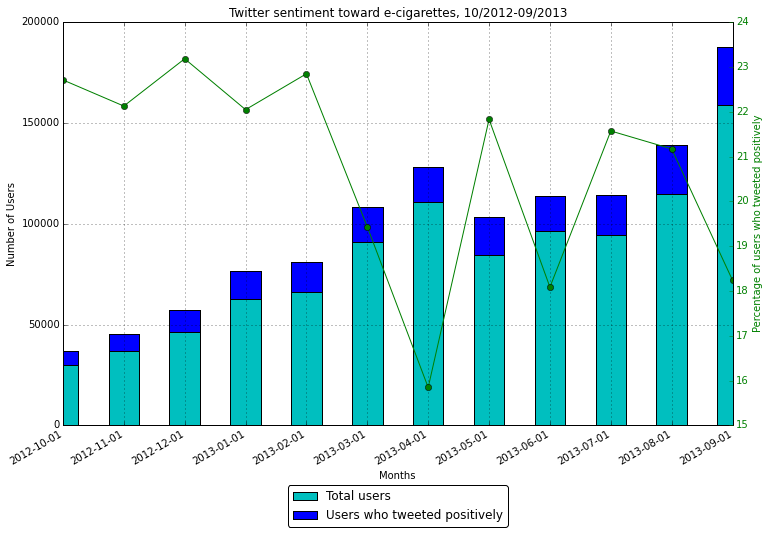


Figure 4 - No RT

With the goal to try to find out why we have those spikes we have analyzed n-grams, which are a sequence of words in the tweet, and the top words for each month



|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| All Organic Tweets | | | | | | | | | | | |
| Oct | Nov | Dec | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep |
| dragonfly 16.81 | election 10.12 | mhealth 9.29 | igotoaschoolwhere 8.52 | whatevs 9.16 | onew 12.53 | courtney 13.21 | efags 10.50 | financile 9.57 | uncut 8.34 | bloomberg 9.17 | whisp 14.50 |
| ltdTHIS\_IS\_A\_URL 8.74 | thanksgiving 8.61 | christmas 8.26 | soar 7.92 | ecigs1 8.46 | 03 10.78 | senators 12.17 | efags2 9.40 | boomed 7.54 | bbj 7.61 | proudtovape 8.25 | worldvapingday 9.76 |
| stoptober 7.89 | election2012 8.17 | youcantbetakenseriouslyif 7.97 | mook 7.03 | valentine 7.69 | pope 9.96 | euecigban 11.55 | uae 8.66 | fathers 7.06 | modeltwo 7.17 | playbook 8.21 | nude 9.41 |
| discos 6.85 | sarapmagvape 7.71 | ganj 7.74 | embarrassment 6.85 | valentines 7.13 | vaporware 8.87 | primarily 9.08 | 1250 8.24 | mhra 6.87 | lirr 7.02 | kv2 7.87 | attorneys 9.05 |
| clik 6.71 | sideburn 6.81 | xmas 7.11 | thingsthatbotherme 6.43 | queers 6.89 | vatican 8.15 | leachon 8.60 | workshop 8.08 | rachet 6.74 | ankle 7.01 | 08 7.61 | patches 8.68 |
| mall 6.56 | pattinson 6.61 | dec 7.00 | 22509 5.97 | superbowl 6.84 | sxsw 8.10 | welcometomyschoolwhere 7.99 | subculture 6.83 | fuels 6.58 | dumbasses 6.57 | smartcigs 7.39 | lab13 8.07 |
| clintandrew 6.29 | shishavapes 5.96 | govawards 6.33 | 36e 5.84 | oscars 6.51 | nanopartices 7.82 | 750 7.90 | imu 6.54 | blucigscoupons 6.56 | manhattan 6.49 | stupider 7.35 | teens 7.88 |
| fuckingsmart 6.28 | nov 5.64 | itsabaddaywhen 6.09 | pst 5.84 | baftas 6.49 | conclave 7.71 | eue 7.79 | jury 6.49 | sowwy 6.48 | muffler 6.09 | mass 7.20 | vapefest 7.61 |
| dispoecig 6.25 | emazin 5.60 | camilla 5.91 | perfects 5.81 | positivity 6.11 | faked 7.49 | mywebcamTHIS\_IS\_A\_MENTION 7.65 | interpretations 6.39 | bonnaroo 6.46 | agrees 5.91 | musicians 6.56 | duluth 7.46 |
| ohmyvapor 6.06 | rob 5.47 | legalise 5.91 | monumental 5.79 | britons 6.00 | quitsmoking2 7.32 | irs 7.57 | blogengage 6.38 | restrictions 6.29 | pnoy 5.78 | ushered 6.51 | students 7.39 |