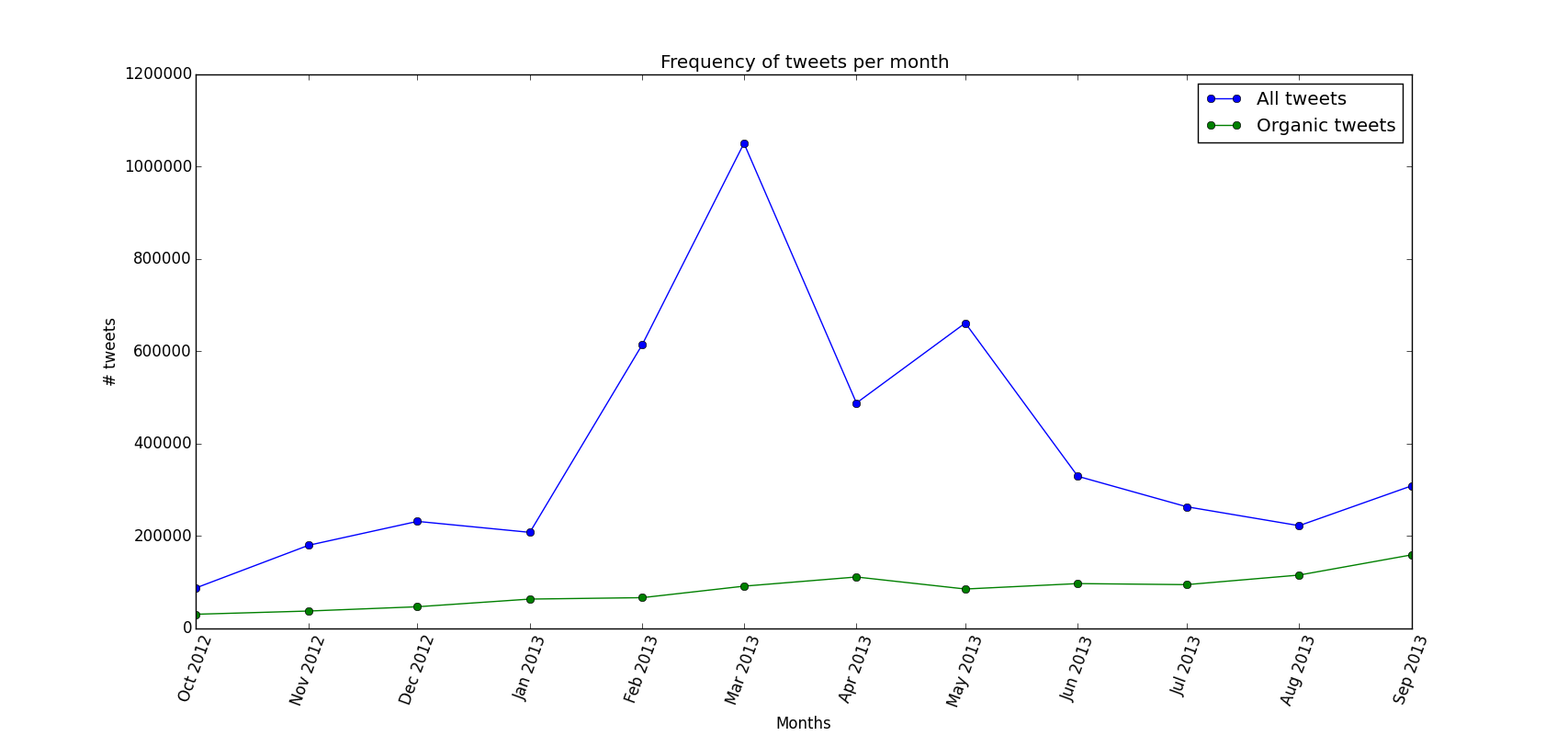
Data

The data is composed for more than 4 million 4639885 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 more than 900k tweets. The plot below shows the number of tweets by month. 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.

Classifier

Training

We have created a classifier using logistic regression, which is responsible to predict the labels of all the organic tweets based in the sample of 2000 we have labeled as training set. We have tested all parameters of this function to have a better accuracy for the classifier and our best classifier has the accuracy of 81.8%.

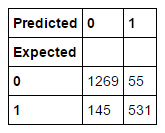


Figure 1 - Confusion Matrix of Logistic Regression Classifier

Precision:[0.89745403, 0.90614334])

Recall:[0.95845921, 0.78550296])

F-score:[0.92695398, 0.84152139])

Testing –

We have used the testing set with all 992k organic tweets. After running the classifier on them we had 27% of positive tweets and 73% of negative.

**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)]

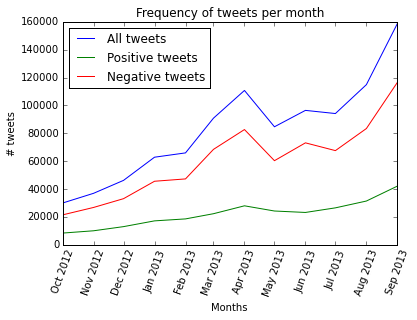


Figure 2 - all organic

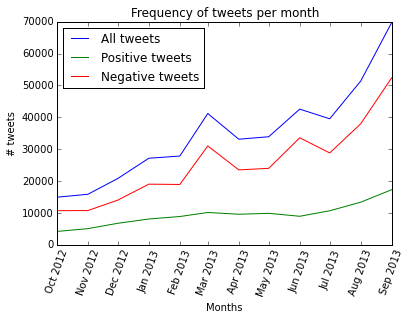


Figure 3 - one tweet per user

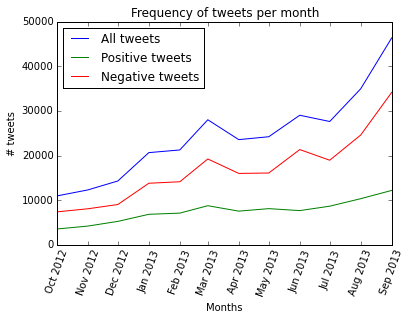


Figure 4 - No RT