ILLINOIS INSTITUTE OF TECHNOLOGY
Departament of Computer Science

Twitter Sentiment Analysis toward e-cigars

Elaine Cristina Resende
Aron Culotta
June 22, 2015

1 INTRODUCTION

In this project we have been working with twitter data which was collected during a year period, from October 2013 through September 2014, looking for mentions and for people talking about electronic cigarettes (e-cigs). The main goal of this project is analyze what Twitter users are talking about this subject, and try to get the sentiment of the users towards e-cigs, in order to find out how people are receiving the use of e-cigs and also to check the type of customers are buying them. (improve it???)

2 TWITTER 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 add up to 992,633 tweets. Figure 2.1 shows the number of tweets (y axis) by month (x axis). The green line refers to organic tweets and blue line to all tweets. We can see that there are spikes in the months of March, May and September (talk about number of organic???).

We have randomly selected 2000 tweets from the set of organic tweets and then we manually labelled those tweets as positive, negative and neutral. For this analysis 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 tweets classified as positive and 1324 as negative.

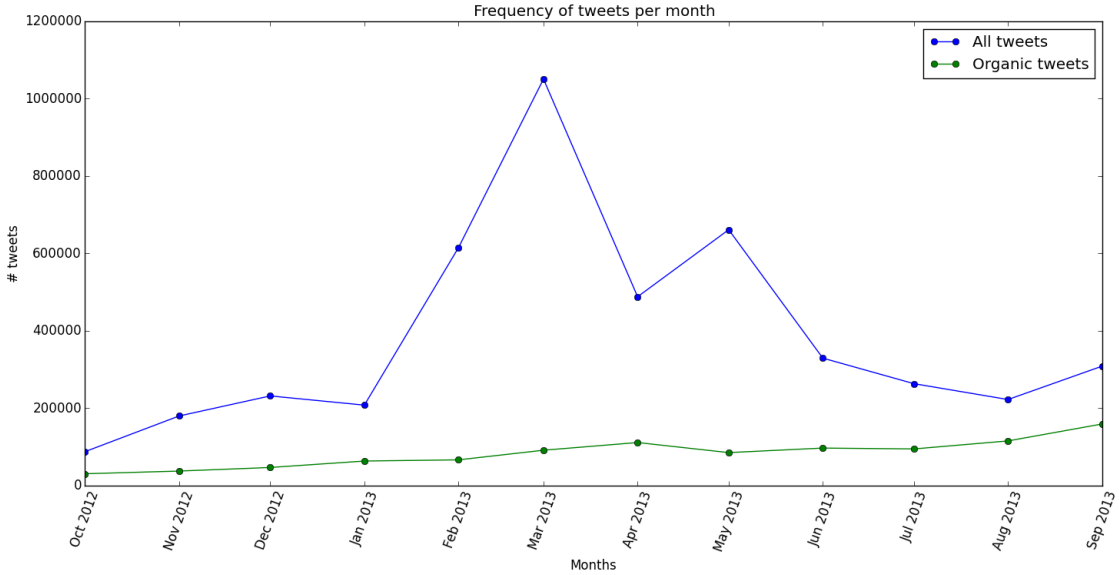

Figure 2.1: Number of Tweets by month

3 CLASSIFICATION TASK

For this task we have built a classifier using Logistic Regression (LG), trained it with the labelled tweets and test it with all organic. Section 3.1 and 3.2 are going to show how was our methodology

3.1 TRAINING

We have created a Logistic Regression (LG) classifier, which is responsible to predict the labels of all the organic tweets. The labelled 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. Figures 3.1 and 3.2 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 3.3 (need to show???). Table 3.1 show precision, recall and F-score values for both classes and as we can notice the percentage of retrieved tweets that are relevant is 90% what is seen as a good value for our assignment.

3.2 TESTING

Our test set is defined as all organic tweets (900k tweets). After running LG 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. Thus, new 200 tweets

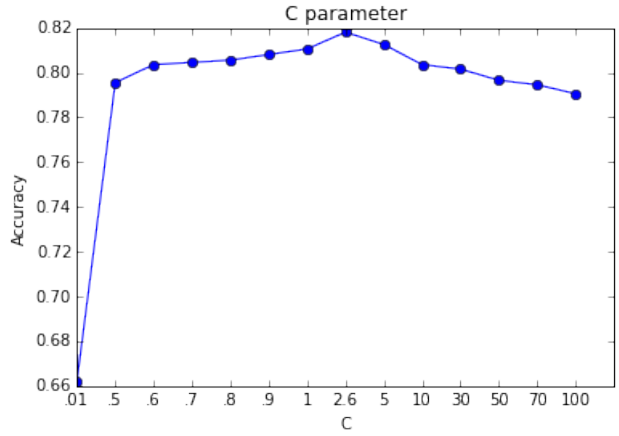

Figure 3.1: C Parameter

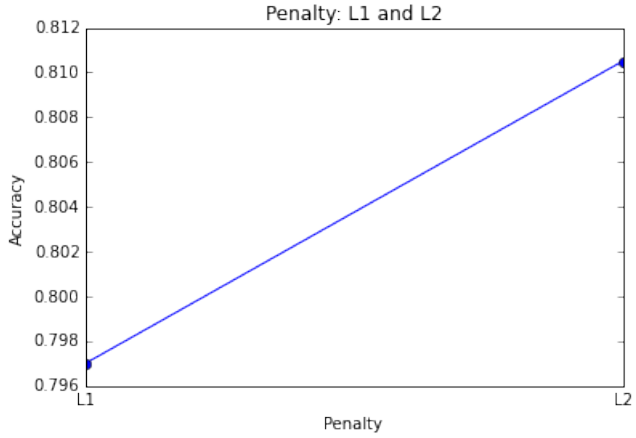

Figure 3.2: Penalty

were randomly selected and manually labelled to be our training set for the checking new LG model. The labels were used as the true labels and the predicted labels were compared to those, in order to see how was our prediction quality. After this model was built, we had 79% of accuracy, so we conclude that our model is good for predicting the sentiment of the tweets (provide better critical analysis???).

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: (check the relevance of this info???if yes, how to show it)

Table 3.1: Classification Quality Measure

Measure	Negative Class	Positive Class
Precision	0.89745403	0.90614334
Recall	0.95845921	0.78550296
F-score	0.92695398	0.84152139

Predicted	0	1
Expected		
0	1269	55
1	145	531

Figure 3.3: Confusion Matrix

4 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 4.1 we have considered all organic tweets, in the Figure 4.2 it is considered just users who tweeted once, in the Figure 4.3 one tweet per user is considered, and in the Figure 4.4 it is not considered re-tweets.

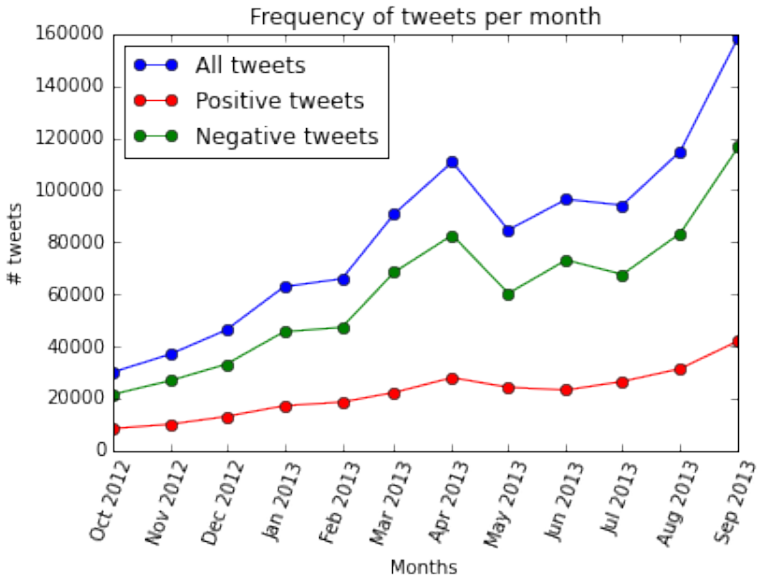

Figure 4.1: All organic

Figure 4.1 shows spikes in April, June and September in all categories, but Figure 4.2, which represents one tweet per user, and Figure 4.3 have moved the first left spike from April to

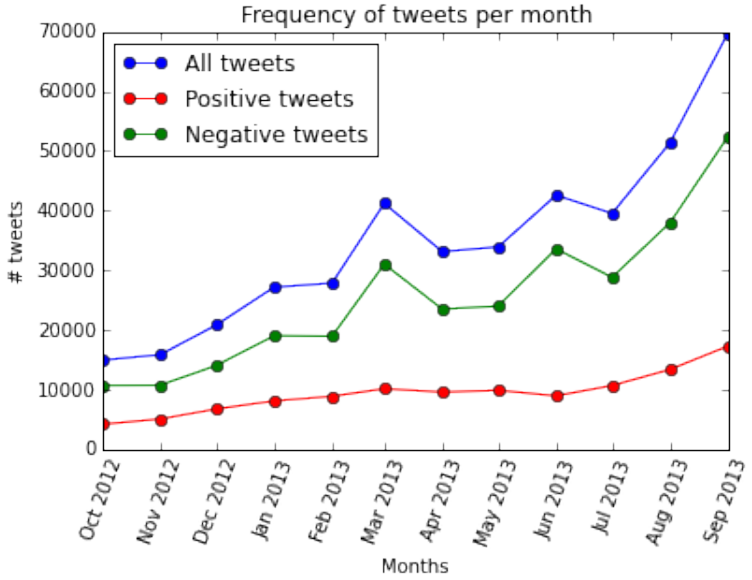

Figure 4.2: Users who tweeted once

March, what tell us that people have posted more tweets in April. Figure 4.2 and Figure 4.4 follow the same shape they differ just in the number of tweets.

Aiming 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 also the top words for each month (shown in the Table 4.1)

March (decide how to show this information... every month or just main ones???)

April

Considering just positive tweets and going through the n-grams of the month of April we found in total 20255 tweets. After analyzing the n-grams, we took a further look in those tweets and we found more than 600 re-tweets of: " @SIMONCOWELL: I HAVE CUT DOWN SMOKING. A LOT. USING ELECTRIC CIGARETTES TO HELP. ARE THEY BAD FOR YOU? ".

According to Figure 4.5 April was the month which people talked more in negative way towards e-cigars. But taking a further look in the Table 4.1, we can see that is because some people were talking about euecigban which was used as a hashtag to fight against the banning of e-cigs [1], some senators in US made regulations to restrict the sale of e-cigars to children [2], and also about an famous TV commercial.[3]

Figure 4.5 shows in the right y axis the percentage of positive tweets over all organic, we can see that the percentage of positivism is higher during December and February an them it goes down during March and April. May starts going up again, but for the rest of the months no one goes upper than May.

[1] https://www.thunderclap.it/projects/7122-stop-the-euecigban

[2] http://www.enewspf.com/latest-news/health-and-fitness/42182-senators-call-on-fda-to-restrict-the-sale-distribution-and-marketing-of-e-cigarettes.html

[3] http://www.huffingtonpost.com/2013/04/03/courtney-love-njoy_n_3006660.html?ncid=edlinkusaolp00000003

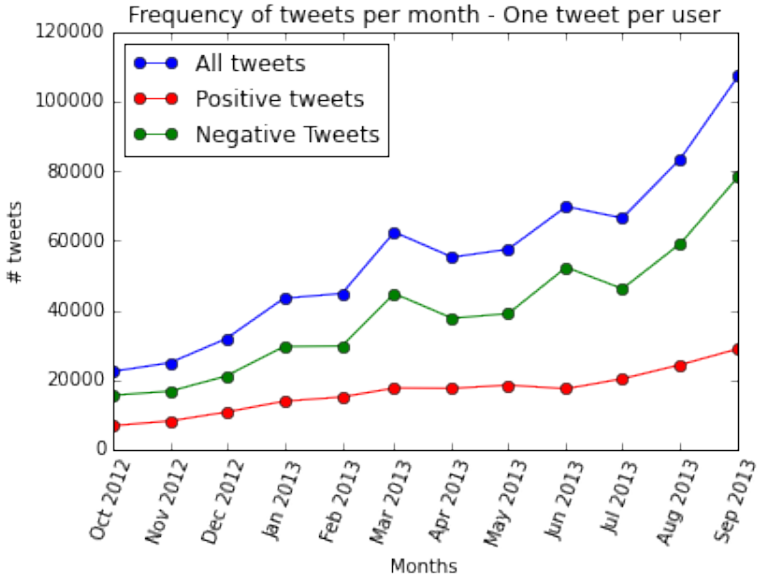

Figure 4.3: One tweet per user

The bars in the Figure 4.5 show the number of users who are posting positively per month in blue color, and we can notice that this number is much less than negative, what tell us that most of people talk negatively about e-cigs. (can we say that???)

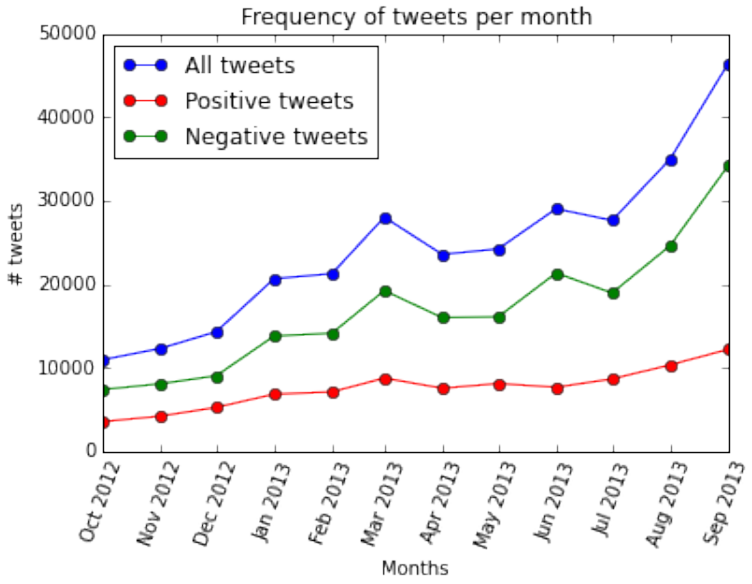

Figure 4.4: No re-tweets

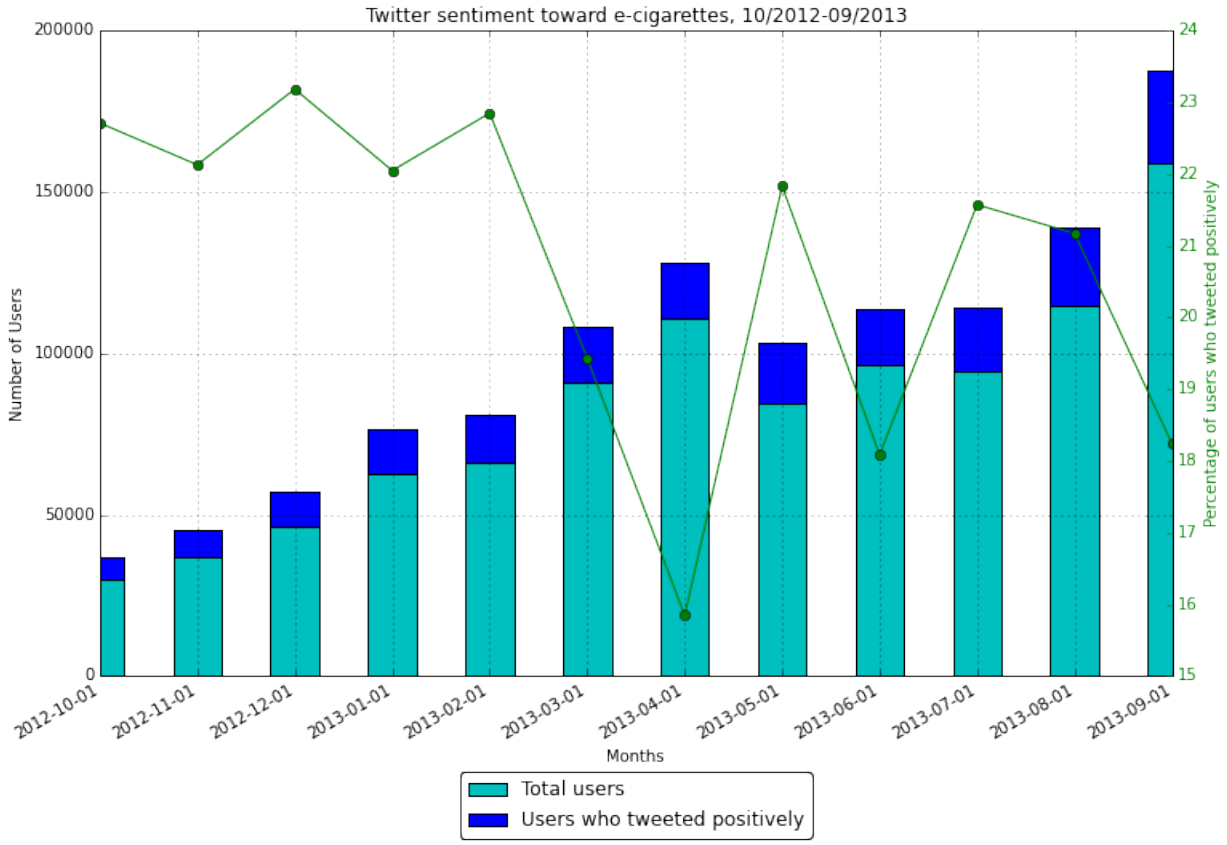

Figure 4.5: Sentiment

Table 4.1: 10 Most significant words by month

Oct	Nov	Dec	Jan	Feb	Mar
dragonfly 16.81	election 10.12	mhealth 9.29	igotoaschoolwhere 8.52	whatevs 9.16	onew 12.53
ltdTHIS_IS_A_URL 8.74	thanksgiving 8.61	christmas 8.26	soar 7.92	ecigs1 8.46	03 10.78
stoptober 7.89	election2012 8.17	youcantbetakenseriouslyif 7.97	mook 7.03	valentine 7.69	pope 9.96
discos 6.85	sarapmagvape 7.71	ganj 7.74	embarrassment 6.85	valentines 7.13	vaporware 8.87
clik 6.71	sideburn 6.81	xmas 7.11	thingsthatbotherme 6.43	queers 6.89	vatican 8.15
mall 6.56	pattinson 6.61	dec 7.00	22509 5.97	superbowl 6.84	sxsw 8.10
clintandrew 6.29	shishavapes 5.96	govawards 6.33	36e 5.84	oscars 6.51	nanopartices 7.82
fuckingsmart 6.28	nov 5.64	itsabaddaywhen 6.09	pst 5.84	baftas 6.49	conclave 7.71
dispoecig 6.25	emazin 5.60	camilla 5.91	perfects 5.81	positivity 6.11	faked 7.49
ohmyvapor 6.06	rob 5.47	legalise 5.91	monumental 5.79	britons 6.00	quitsmoking2 7.32

Apr	May	Jun	Jul	Aug	Sep
courtney 13.21	efags 10.50	financile 9.57	uncut 8.34	bloomberg 9.17	whisp 14.50
senators 12.17	efags2 9.40	boomed 7.54	bbj 7.61	proudtovape 8.25	worldvapingday 9.76
euecigban 11.55	uae 8.66	fathers 7.06	modeltwo 7.17	playbook 8.21	nude 9.41
primarily 9.08	1250 8.24	mhra 6.87	lirr 7.02	kv2 7.87	attorneys 9.05
leachon 8.60	workshop 8.08	rachet 6.74	ankle 7.01	08 7.61	patches 8.68
welcometomyschoolwhere 7.99	subculture 6.83	fuels 6.58	dumbasses 6.57	smartcigs 7.39	lab13 8.07
750 7.90	imu 6.54	blucigscoupons 6.56	manhattan 6.49	stupider 7.35	teens 7.88
eue 7.79	jury 6.49	sowwy 6.48	muffler 6.09	mass 7.20	vapefest 7.61
mywebcamTHIS_IS_A_MENTION 7.65	interpretations 6.39	bonnaroo 6.46	agrees 5.91	musicians 6.56	duluth 7.46
irs 7.57	blogengage 6.38	restrictions 6.29	pnoy 5.78	ushered 6.51	students 7.39