Mining the Propagandistic Web: is Twitter the New Propagandistic Tool in Spain?

Pablo Alonso Baldonedo 463689

May 15, 2015

1 Introduction and motivation

Since the origin of the art of politics, propaganda has been always a core part of it. Being able to gain support from the big mass has been a key factor for revolutions and the road to power. According to Niccolo Maquiavelo *"Every one sees what you appear to be, few really know what you are, and those few dare not oppose themselves to the opinion of the many, who have the majesty of the state to defend them"*. For this reason having the power to build the public appearance of oneself is the golden dream of many politicians, and the way of doing this is through propaganda. But not only politicians, who aim to win elections, but also revolutions can be supported by propaganda. For Vladmir Ilich Ulinov, better known as Lenin, "Lies are revolutionary weapons" and "If you repeat a lie often enough, people will believe it" (sometimes this quote is attributed to Joseph Goebbels).

Propaganda techniques have evolved over time. Rhetoric was created by Greeks and Romans used originally in a series of litigation in ancient greek regarding expropiated citizens. With the end of the *Ancien Régime* and the beginning of the modern democracies propaganda seems to be a central part of our daily lives, specially every four or five years when is time of elections. First the press, then the radio and finally the television, all caused huge changes on how propaganda was made. In the recent years a new mass media has appeared: the internet. Internet changes completely the way the media communicates and adds the possibility of interaction with the public. Social networks are one of the most important part of the internet regarding communication and it has been seen their impact in past elections such as the crowd-funding campaign of Barack Obama in his run for president of United States.

In Spain, huge political changes are going on and the public opinion is changing rapidly. Every month we can read on the press more and more political surveys with surprising results showing the end of a traditional two-party predominance in spanish politics. Two parties PODEMOS and Ciudadanos (C's) seem that are going to burst into the parliament very strongly [1], [2]. PODEMOS has been the political surprise in the last elections for european parliament in 2014. A party that is hardly one year old got over one million votes. Leaving aside the social situation and political analysis of spanish situation, topics completely out of the scope of this project, it is striking that a completely unknown party gets the trust of so many people. In the last year spanish society could see how a party lead by a charismatic leader, famous in political tv shows, was constantly being trending topic on twitter, a microblogging social network. It was the first

time that in Spain a political party appeared massively in social networks and used them in order to gain popularity. The party has put a huge effort and human resources [3] to have a good image in social network.

Trending topics are hot topics on twitter. However, it seems strange that a new small party such as PODEMOS can be often trending topic spontaneously. Our hypothesis is that a small group of people, related to the party, is organized in order to use twitter, and social networks and the internet in general, as a new propagandistic tool. In this project we analyze the content generated under some trending topics related to PODEMOS and try to find such group of users and understand their behaviour.

2 Collecting the data

We have used the twitter API [4] to collect the tweets of our data-set. We have manually chosen several Trending Topic hashtags that were related to PODEMOS, either created by the party itself, either related to a TV political show were PODEMOS was the main topic. We encounter a major problem during this process as twitter API does not allow to search tweets within a hashtag older than 1 week, which was the case for most of our hashtags. To overcome this problem we use the website Topsy [5] which has collected all tweets since 2006. Unfortunately, this service is not free. However, topsy website has around a thousand tweets openly available. We built a crawler that dealt with javascript generated websites like Topsy to retrieve those tweets ids. Once we have collected tweet ids from Topsy we have requested the tweets information one by one through the twitter API, as it is possible to retrieve tweets older than a week if we request them by their tweet id. We have also asked for the retweets derived from each of our tweets. The whole process was quite slow taking around 4-5 hours to mine just one hashtag.

The tweets collected from Topsy may be repeated and tweets and retweets are mixed. Therefore, we conducted a cleaning step were we split our dataset in two: tweets and retweets. We also made sure that there were no repeated tweets in the dataset. In table 1 we show the final results for our dataset for the 17 hashtags that we collected. Note that there are a few that contains many more tweets. This is the case because we were able to track those hashtag realtime with twitter API.

We need a baseline to be able to make some judgements about our results. For this reason we have build a "control dataset" formed by tweets and retweets related to La Liga BBVA and european football in general. We

Hashtag	Tweets	Retweets	Total
#AsambleaCiudadana	433	469	902
#CrisisBipartidismoM4	455	1173	1628
#ElCambioEmpiezaEnAndalucia	351	2001	2352
#L6Nmarchapodemos	490	2487	2977
#MarchaCambioPodemosM4	295	1067	1362
#NoOlvidamosPPSOE	157	1949	2106
#PabloIglesiasResponde	374	973	1347
#Podemos25M	215	817	1032
#PodemosALaGriegaM4	391	1628	2019
#PodemosMarchaARV	388	887	1275
#Razones31E	448	1292	1740
CocinaCasera	155	504	659
IgualesPodemos	4173	12494	16667
Podemos22M	6152	36306	42458
SoloConTuAyuda	1965	5255	7220
VotaPodemosAndalucia	13806	62383	76189
YoVoy20M	5380	21127	26507

Table 1: Dataset information

have chosen this topic because to make a sensible comparisons we need a topic that can generate periodic and different trending topics. However, due to time limitations, we have tracked only hashtags of one weekend. This does not generate problems in the amount of information, as the number of trending topics is comparable to the one we have in our dataset and the number of tweets seems enough. However, it seems that the dataset could be more representative if it was made out of several weekends. This "control dataset" will be used in the rest of this project to make comparisons to our dataset to try to find distinct behaviours in it.

3 Analyzing the dataset

Before carrying any experiments on the dataset we decided to have a look to the dataset. We have already discussed in the previous section and in table 1 how unbalanced in number of tweets and retweets are our hashtags due to the twitter API time limitations. In Figure 1 we show the proportion of tweets and retweets per hashtag. We can see that most of the information

comes from retweets.

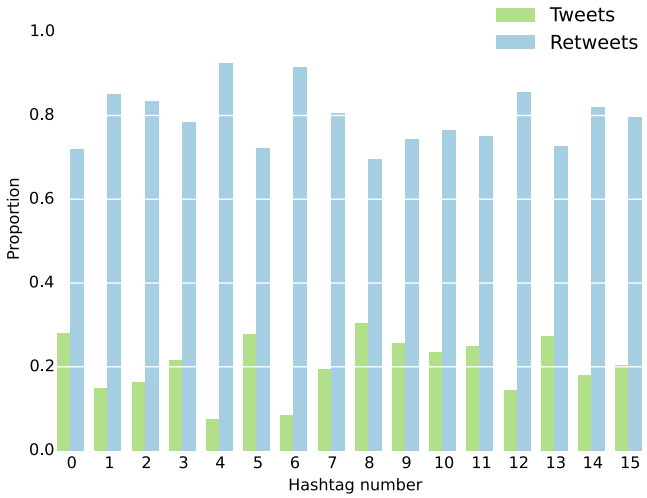


Figure 1: Proportion of tweets and retweets for each hashtag in the dataset.

We have conducted the same study in our "control dataset" as shown in Figure 2. We can see that the proportions look very differently in this case. It also disagrees with what is reported in [6] where it is said that trending topics have around 31% of retweet rate. This means that our dataset, due to the way it was collected through Topsy, is retweet biased. This can distort our results. However, we could not track real time more hashtags (which would overcome this problem), therefore, we have to assumed this bias in our dataset.

4 Clustering

Once we got the data in a suitable format we sought to cluster the users in groups. The idea is to find a group of users that can be labeled as "propagandists", i.e. people who are trying to push a PODEMOS related hashtag into trending topic or, once it has reached that category, keeping it as such. We tried several clusters numbers and different set of features. Finally, we worked with two features: number of tweets in our dataset and number of tweets per hashtag. The idea is to trying to capture users that can either tweet a lot on PODEMOS topic or, at least, have a good rate

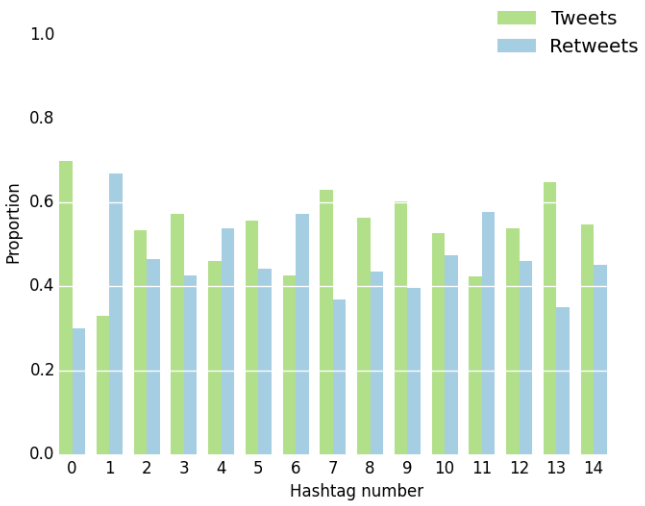


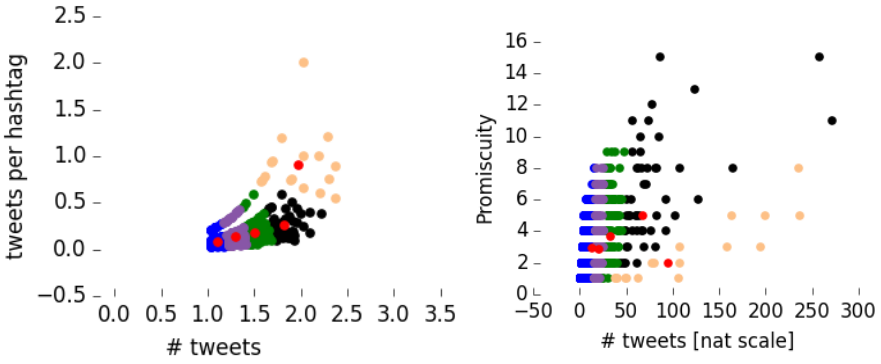
Figure 2: Proportion of tweets and retweets for each hashtag in the "control dataset".

tweets per hashtag.

We have used k-means as clustering algorithm with k++ initialization. The main drawback of this algorithm, apart of the possibility that it gets stuck in local minima, is that we need to set the number of clusters. This procedure is not trivia. We tried different number of clusters and ended up with k=4. We have performed the clustering analysis for both retweets and tweets independently.

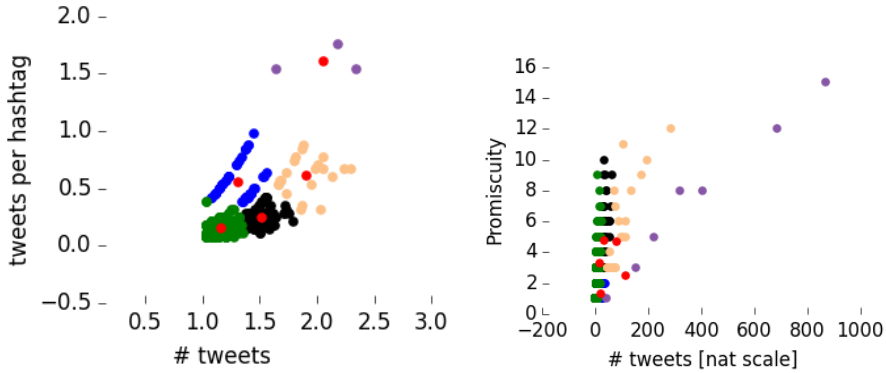
In Figures 3 and 4 we show the results. For avoiding some problems that normally arise when using k-means we have scaled the input variables and removed some outliers. Another problem that we encountered was the fact that our hashtags were very unbalanced in number of tweets and retweets. This could lead that only two or three hashtags that contained many more tweets could determined the whole analysis. Therefore, we have clipped the number of tweets and retweets to a maximum of 3000 per hashtag so that all hashtags are comparable.

We have done the same experiment with our "control dataset". The results are shown in Figures 5 and 6.



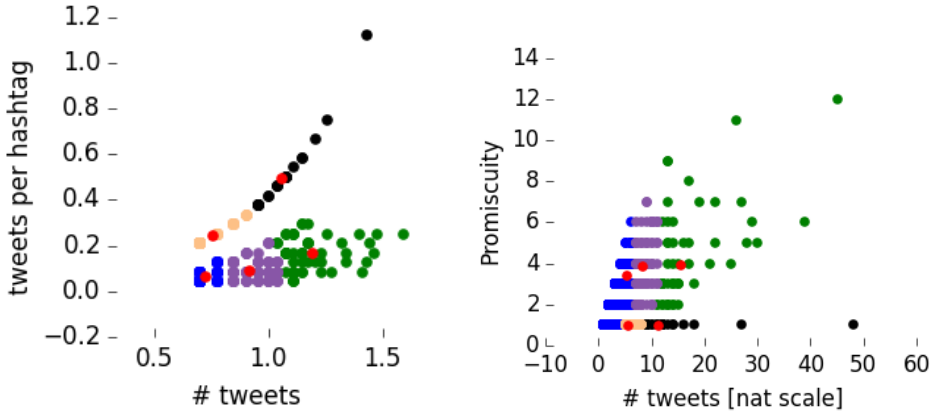
(a) (b) Clusters

Figure 3: Clustering of the retweets. (a) Input to the K-means algorithm with outliers removal and scaled components. (b) Resulting clusters in the space with features "number of retweets" and "promiscuity", i.e, number of different hashtags the user has retweeted.



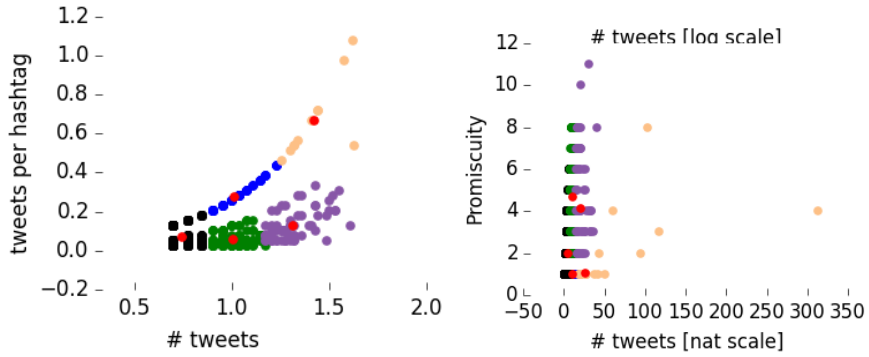
(a) (b)

Figure 4: Clustering of the tweets. (a) Input to the K-means algorithm with outliers removal and scaled components. (b) Resulting clusters in the space with features "number of tweets" and "promiscuity", i.e, number of different hashtags the user has tweeted.



(a) (b)

Figure 5: Clustering of the retweets for the "control dataset". (a) Input to the K-means algorithm with outliers removal and scaled components. (b) Resulting clusters in the space with features "number of retweets" and "promiscuity", i.e, number of different hashtags the user has retweeted.



(a) (b)

Figure 6: Clustering of the tweets for the "control dataset". (a) Input to the K-means algorithm with outliers removal and scaled components. (b) Resulting clusters in the space with features "number of tweets" and "promiscuity", i.e, number of different hashtags the user has tweeted.

5 Propagandist information

We have merged some of the clusters shown in the previous section into a bigger one which forms the "propagandist" cluster. We have ended up with 131 people in such a cluster (1.32% of total users in the dataset). 34 out of the 131 were PODEMOS official accounts. However, most of them do not show any direct affiliation with PODEMOS. In Figure 7 we show some profile photographs of these users. As we can see there is nothing that reminds PODEMOS in them. If we have a look to their user description we get something similar. Some may show some political condition like "Anticapitalist", "against the elites", "Fight against corruption" etc. A few, do have messages that explicitly mention their support to PODEMOS, mostly with puns like "Confio y deseo que todos juntos PODEMOS cambiar las cosas" (I trust and wish that all togheter we CAN change things).

We found some interesting results. Our algorithm classified as "propagandist" the official account of the TV show "La Sexta Noche" which is probably the most influential political show this day in Spain. This may make sense as the official account of a TV will try to make the hashtag of its program trending topic so will behave as a propagandist. We also found a user that, actually, is always criticizing PODEMOS. It seems that is a person that engages a lot the political discussion in twitter and acts as a "counter propagandist" so it is reasonable that the algorithm classifies in this cluster. This will also lead to a further discussion such as if a sort of sentiment analysis will be useful as an extension of this work for performing the clustering.

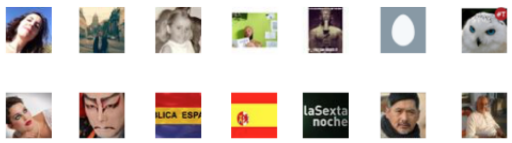


Figure 7: Profile photographs of some propagandists.

6 Propagandist activity in hashtag evolution

We have studied which is the influence of these "propagandist" in the hashtags over time. In Figures 8, 9, 10 and 11 we see the proportion of tweets, retweets and tweets + retweets generated by these propagandists in some

hashtags over time. We can see that in the first two a large of the proportion of the information generated under the hasthag comes from this group. However, in the two last ones the proportion is quite low, not even reaching 10 %. This is, however, reasonable as the last two corresponds to TV shows so the topic can become hashtag by itself or at least does not require that much help from propagandists to become so.

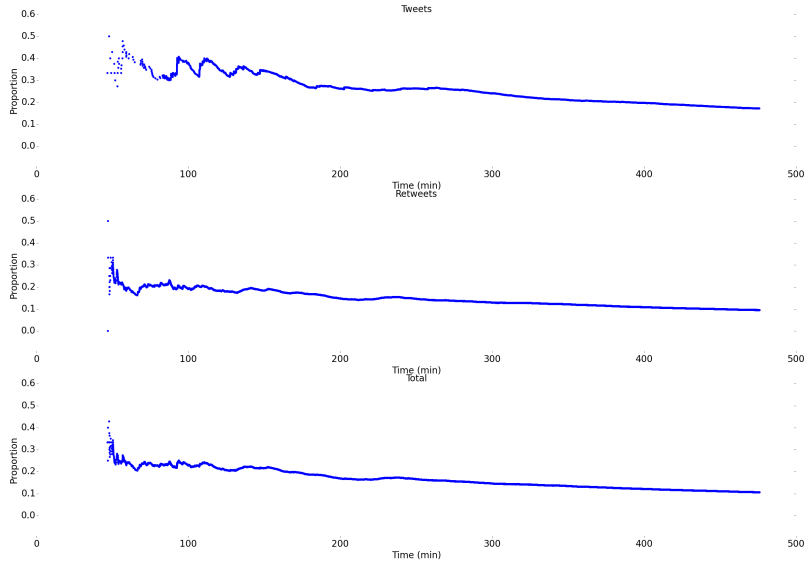


Figure 8: Proportion of tweets (up), retweets (center) and tweets + retweets (bottom) made by the propagandist over time under the topic Podemos22M.

The next question we asked ourselves was if this proportions are relevant. For this reason we have conducted the same study in our "control dataset". Most of the hashtags has a "propagandistic" engagement level below 10% in all tweets, retweets and total. There are a few exceptions like the ones shown in the Figures 12, 13 and 14. However most of the hashtag followed a trend like the one in Figure 15. Therefore, we can say that most of the "control" dataset hashtag has a completely different behaviour than the one observed in our political dataset. Only "Ancelotti" and "Carvajal" topics seems a little bit closer in behaviour, however, in none of them we reach levels of engagement over 30% or even 50% as we can see in our political dataset. Also "Simeone" shows a relatively high engagement but it soon

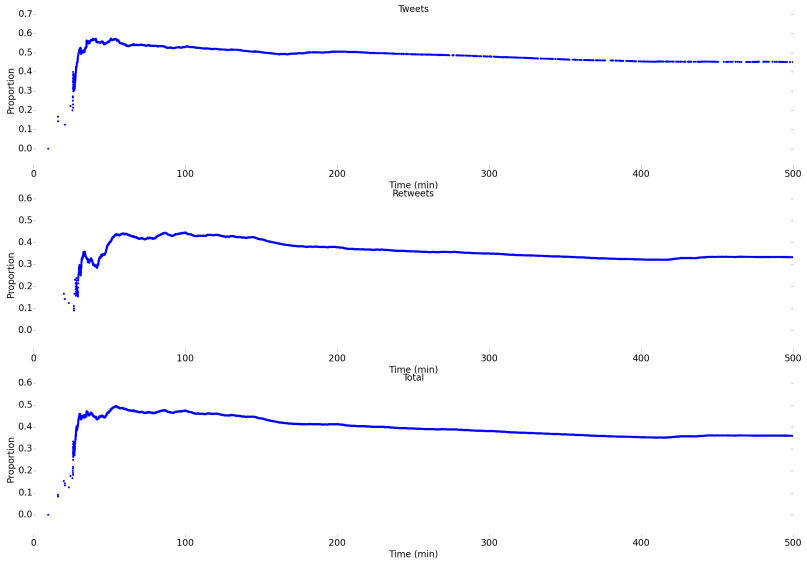


Figure 9: Proportion of tweets (up), retweets (center) and tweets + retweets (bottom) made by the propagandist over time under the topic YoVoy20M.

declines.

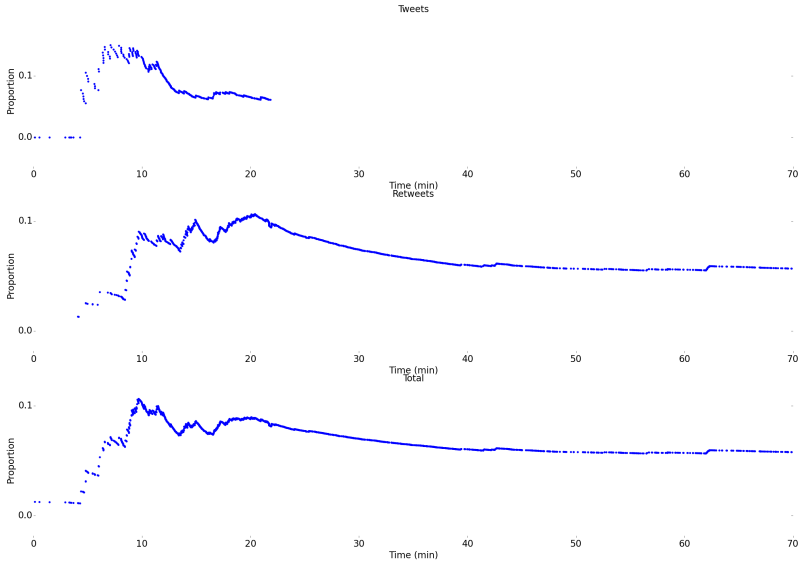


Figure 10: Proportion of tweets (up), retweets (center) and tweets + retweets (bottom) made by the propagandist over time under the topic #L6Nmarchapodemos.

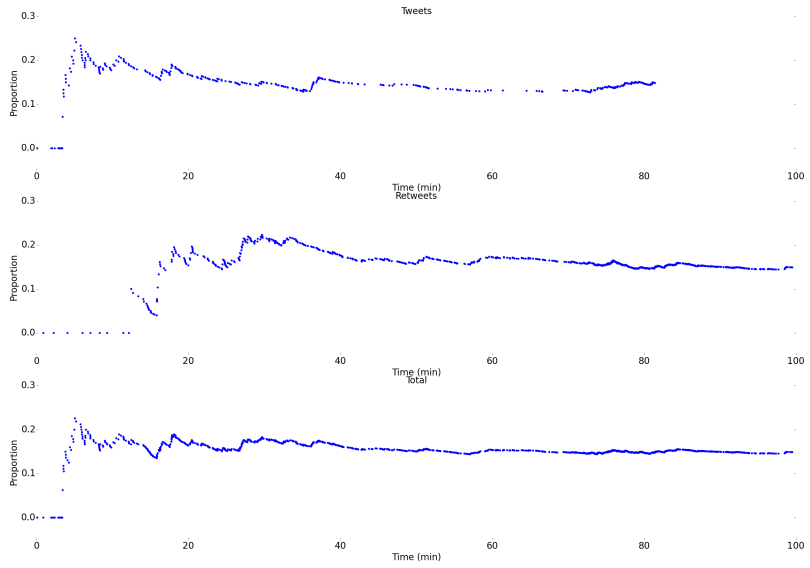


Figure 11: Proportion of tweets (up), retweets (center) and tweets + retweets (bottom) made by the propagandist over time under the topic #Podemos-ALaGriegaM4.

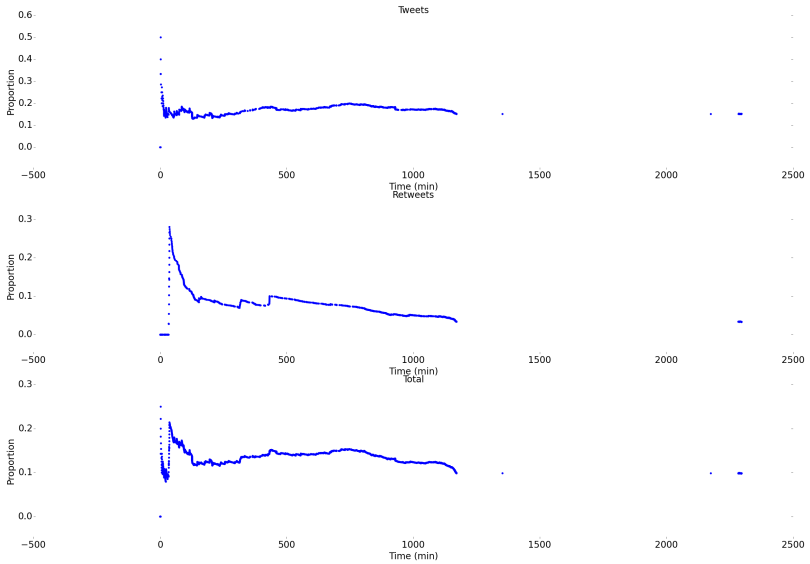


Figure 12: Proportion of tweets (up), retweets (center) and tweets + retweets (bottom) made by the propagandist over time under the topic Ancelotti.

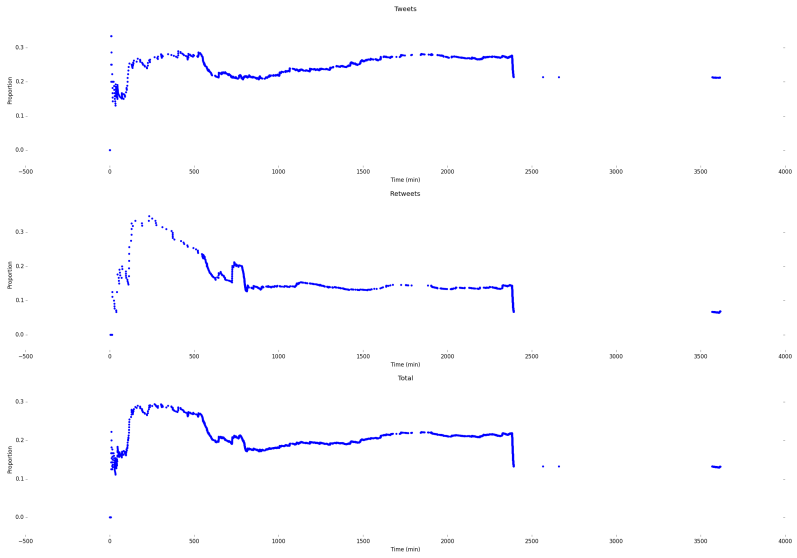


Figure 13: Proportion of tweets (up), retweets (center) and tweets + retweets (bottom) made by the propagandist over time under the topic Carvajal.

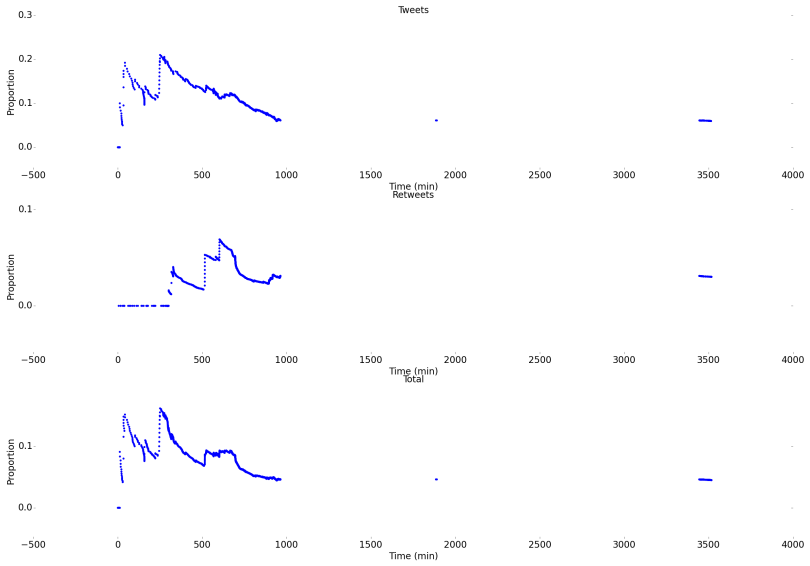


Figure 14: Proportion of tweets (up), retweets (center) and tweets + retweets (bottom) made by the propagandist over time under the topic Simeone.

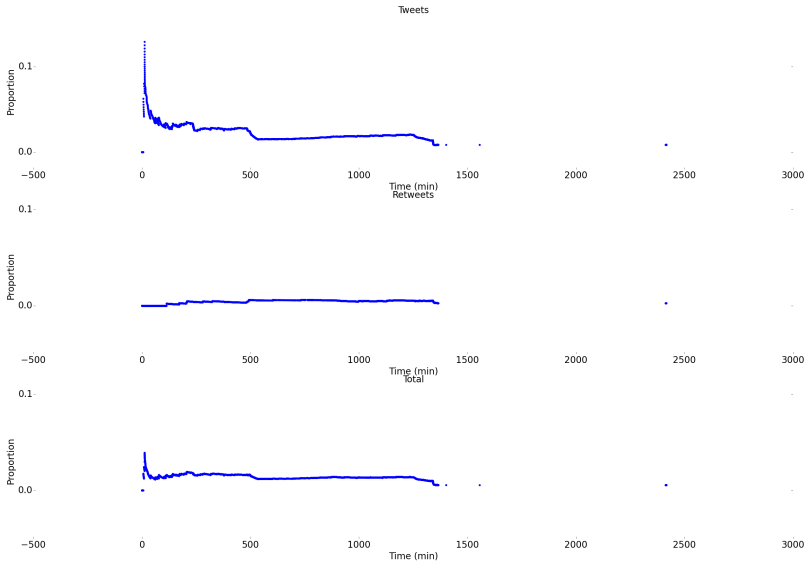


Figure 15: Proportion of tweets (up), retweets (center) and tweets + retweets (bottom) made by the propagandist over time under the topic Chicharito.

7 Friendship graphs

We have explored who are these propagandists following. In Figure 16 we show the friendship graph, i.e., our *propagandistic* users and the people they are following. In the graph the propagandists are shown in blue and the other users in red. We can see that the users that are being followed by most of the propagandists (bigger nodes) are mainly outside the propagandistic group. In Figure 17 we show the same graph but after performing some filtering. Now only users who are being followed by at least 60 propagandists are shown and we have zoomed the part of the graph were the most popular users lie. These users are mostly PODEMOS official and leaders accounts. There are also some other politicians from leftist parties like *@agarzon* (leader of "Izquierda Unida", "United Left") or *@ainhat* (former member of "Izquierda Unida" and former Pablo Iglesias, PODEMOS leader, girlfriend). Some journalists such as *@_anapastor_* or *@jordievole* also appear along with some (left oriented) media such as *@eldiarioes*, *@publico_es* or *@la_tuerka* (Pablo Iglesias' own TV political show).

Figure 16: Friendship graph. In red *propagandistic users* and in blue the rest. The size of the node is proportional to the number of *propagandistic* followers the user has.

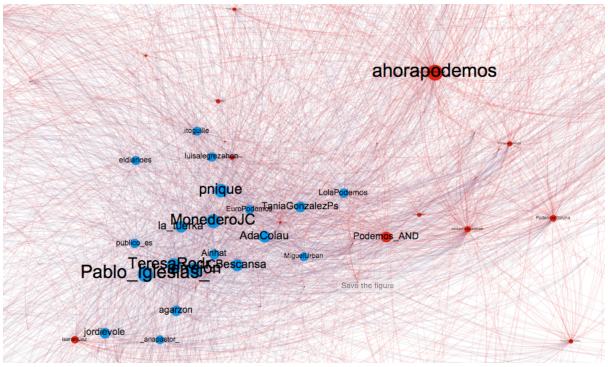


Figure 17: Friendship graph when considering users that are being followed by at least 60 propagandists. In red *propagandistic users* and in blue the rest. The size of the node is proportional to the number of *propagandistic* followers the user has.

8 Retweet graph

We are now interested in who are the people that generates the content that the propagandists retweet. In Figure 18 we show precisely this. Again red nodes are propagandists and in blue the other users. In Figure 19 we filter the graph for users being retweet at least 30 times by propagandists. We can see that, as in the friendship graph, the most retweeted users are not propagandists and corresponds to Podemos offical accounts or Podemos leaders.

Figure 18: Retweet graph. In red *propagandistic users* and in blue the rest. The size of the node is proportional to the number of *propagandistic* retweets.

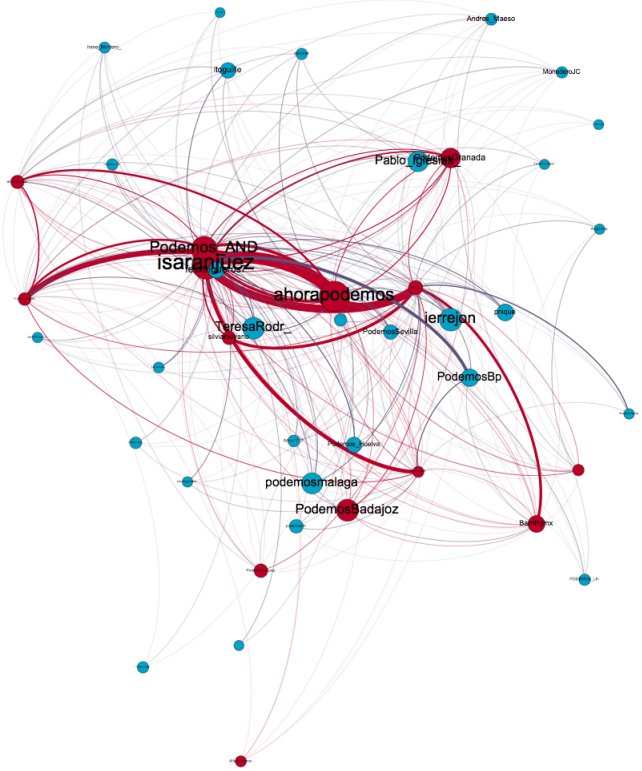


Figure 19: Retweet graph when considering users that are being retweet at least 40 times by propagandists in our dataset. In red *propagandistic users* and in blue the rest. The size of the node is proportional to the number of *propagandistic* retweets.

9 Propagandist behaviour

We have studied the tweeting and retweeting behaviour of the propagandists and check if they act differently than the rest of the users. In Figure 20 we show how long does it take, in median, to the propagandist to retweet from the moment the original tweet was posted. We can see that half of total propagandists retweet before 10 minutes from the moment the original tweet was posted. In Figure 21 we see that, when taking all the users, the behaviour is quite different.

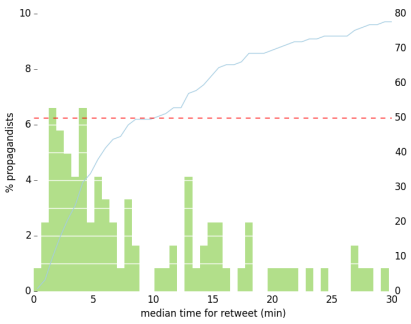


Figure 20: Median time elapsed before a propagandist makes a retweet from the moment the original tweet was posted. In blue the cumulative number of users (scale in the right axis).

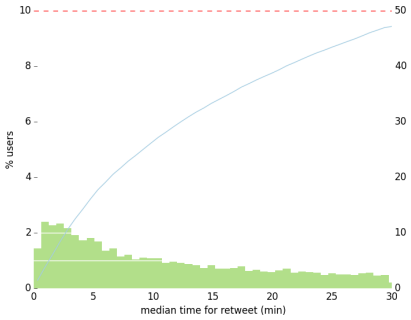


Figure 21: Median time elapsed before a user makes a retweet from the moment the original tweet was posted. In blue the cumulative number of users (scale in the right axis).

We have conducted the same experiment with the tweets analyzing the

time from the first tweet under the hashtag. In Figure 22 we show the results for the propagandists and in Figure 23 for all users. We see again a very different behaviour, being the propagandists quite faster.

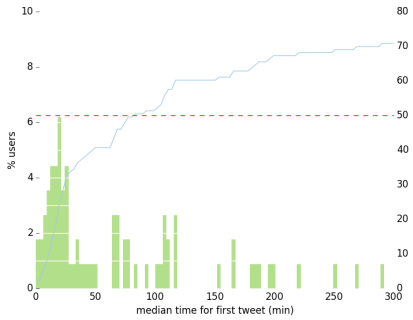


Figure 22: Median time elapsed for a propagandist to tweet under a hashtag since the first tweet was posted. In blue the cumulative number of users (scale in the right axis).

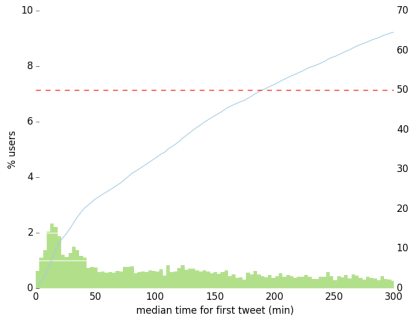


Figure 23: Median time elapsed for a user to tweet under a hashtag since the first tweet was posted In blue the cumulative number of users (scale in the right axis).

10 Conclusions

We have analyzed a dataset built from tweets related to the spanish political party PODEMOS under the hypothesis that there exist a group of people, that we have named “propagandists”, who act as a group and try to make

trending topic PODEMOS related hashtags. We have cluster around 1% of the total users as "propagandists". We have seen how these users created a significant proportion of the total content on the hashtags, which is something not common as seen in our "control dataset". On top of that, their behaviour of tweeting and retweeting is significantly different with a tendency of engaging the conversation in the early stages of the hashtag and with a faster retweet time delay than the rest of the users. For these reasons we can claim that our hypothesis was true and there exist a group of people whose goal is to make popular topics of their interest.

From the graphs we can say that the content-makers are not member of the "propagandist" group but rather PODEMOS leaders and official accounts. The "propagandists" tend to retweet and expand those contents but do not create as much content as the one they retweet.

Finally, we have to bear in mind that our database was quite limited in number of tweets and somewhat biased towards retweets. Therefore, our claims should be taken with caution as these facts may have distorted our results. The question is if the limited amount of tweets that we worked with are a representative sample of the real population of tweets which is much bigger. As future work, we should save tweets real time to have a good dataset and redo the whole experiments again to confirm or dismiss the claims made in this project.

References

[1] http://www.wsj.com/articles/spains-new-leftists-face-first-domestic-election-1426791383

[2] http://www.wsj.com/articles/with-plan-to-fix-spains-economy-political-upstart-surges-in-popularity-1429186998

[3] http://vozpopuli.com/actualidad/58270-el-ejercito-de-podemos-en-twitter-mas-de-20-soldados-para-gestionar-la-cuenta-oficial-del-partido

[4] Twitter REST API: https://dev.twitter.com/rest/public

[5] www.topsy.com

[6] Sitaram Asur, Bernardo A. Huberman, Gabor Szabo and Chunyan Wang. Trends in Social Media : Persistence and Decay.

11 Appendix

Code project can be found in github pabaldonedo/truth-minister. The datasets are not available due to space limitations in github. However, I will provide it to whoever that asks for it.