Analysis of Twitter followers of leading international companies.

Quantitative and qualitative study of behaviours demonstrated by humans

(users which are presumably real) or by bots (users which are presumably fake).

Author:
Marco Camisani Calzolari – Teaching Professor in Corporate Communication and Digital Languages - IULM University of Milan
marco@camisanicalzolari.com
Telephone+39-02-303123179
www.camisanicalzolari.com

8th June 2012

Abstract

Twitter is a free service which enables its users to update their profile through text messages of up to 140 characters in length. It has reached 500 million users, of which only 140 million are active users (who have posted a tweet in the past 30 days). Each user can "follow" other users, and be followed by others in turn. A user's followers (other users following him or her) will receive all his messages on their homepage, and likewise he/she can follow other users.

The vast number of registered users has turned Twitter into a marketing tool used by companies to promote their products. Many businesses have created a corporate account in order to promote themselves with the aim to gathering as many followers as possible.

The ultimate objective of marketing activities, including digital marketing, is to sell more products. However, the number of followers is still a marker which is often perceived as being representative of a company's success on Twitter, regardless of the real value of its followers in terms of interest in the brand or in its products.

Comparisons with their competitors, in terms of number of followers, has led some companies, whether directly or indirectly through intermediaries, to purchase followers for the sole purpose of increasing the overall number. The high demand has created an international market for selling followers, real or otherwise.

But what is a real follower as opposed to a fake one? We can consider as being real those users who correspond to actual individuals, who through their own Twitter account decide to voluntarily follow the company's profile.

The market offers totally fake followers, in other words created by bots, i.e. robot software programs capable of generating vast quantities of accounts created ad hoc, which then become the followers of those willing to pay for them. It is not possible to distinguish a fake follower from a real one with utter certainty; however, it is possible to identify characteristics and behaviours that are typical of real users and of bot users. The results are based on an algorithm that is capable of assigning "human behaviour" points and "bot behaviour" points using a software created especially for this purpose. The method and the algorithm were defined according to criteria which I personally believe to be well-founded and credible, but it should be pointed out that other researchers many apply different values to the parameters or use other methods, thus arriving at different results. In this study, I have taken into account the main international brands, starting with a sample of their followers, and I have given each a value which indicates how many of their followers demonstrate human or, conversely, bot behaviours. The figures which have emerged show that there are large numbers of what are most probably bots, or at least inactive users, following the profiles of major brands on Twitter.

The algorithm distinguishes between active users and inactive users, and as far as reasonably possible, between those which are most likely "human" and those which are probably "bots". It is not possible to state with absolute certainty whether the users which "bot"-identified behaviours have been artificially generated or instead correspond to human users. However, the results show a remarkable difference in terms of the number of potential bots between one brand and another. To some extent, this confirms the precision of the algorithm, as otherwise it would have produced similar results without marked differences between the various brands.

Foreword

This study offers an overview of the current situation of the most important companies on Twitter. It provides a basis for weighing the real value of the number of followers of companies, which is often seen as representing how successful they are on Twitter.
No case studies are available which illustrate the entire panorama of brands on Twitter; there are only case histories of certain successful brands which have actually found Twitter to be an effective way of offering customer service or engagement.
This study fills part of that void, offering some necessary groundwork for analysing the behaviour displayed by the Twitter audiences of the most important corporate profiles, which may be followed up with qualitative research.
This study is innovative in that it has processed the characteristics of a large number of users in order to highlight those profiles which act in ways that differ from normal human behaviour on social networks.

Experimental Conditions

The research was carried out using software designed especially by The Fool, a company which specialises in creating forensic software programs for analysing user behaviours on the Internet. In practice, they provided a program which is capable of analysing a random sample of the Twitter followers of the companies I examined; using objective parameters, the program reported on which characteristics or behaviours can be attributed to a human rather than a bot.
The software extracted the data in .csv format. These files, one for each company, contain data under each of the following headings:

> Father_ScreenName: The Brand account's Twitter name
> Father_Following: The number of Twitter accounts that the Brand is following
> Father_Followers: The number of followers that the Brand has
> Follower_Id: The Twitter_id of the user in the Panel
> Follower_ScreenName: The Twitter name of the user in the Panel
> Follower_Protected: If TRUE the user in the Panel is private and does not have any statistics
> Follower_HumanValue: The Human Points of the user in the Panel
> Follower_BotValue: The Bot Points of the user in the Panel

This data was aggregated and inserted into a Mysql database from which, using a composite query, we extracted the overall results, which were then exported to Excel worksheets so they could be reordered and compared.

Sample

Basic criteria: the study only examined corporate accounts, and only those of companies selling products or services. Local, personal, or celebrity accounts were excluded, as were those of media companies such as radio, television or newspapers.
Based on the aforementioned filters, I defined three groups of accounts:

· **International companies in the world**
· **International companies in Italy**
· **Italian companies**

On May 20th 2012, the companies were selected from the following sources:

· **International companies in the world**

From the top 1000 (by number of followers) from http://twitaholic.com/, only those companies corresponding to the aforementioned basic criteria were extrapolated.
These numbered only **13** in all.

· **International companies in Italy**

From the top 1000 (by number of followers) from http://twitaholic.com/, only those companies indicating Italy as their location and corresponding to the aforementioned basic criteria were extrapolated.
These numbered only **6** in all.

· **Italian companies**

These were identified using two different sources:

The first was http://twittercounter.com/pages/100?time_zone=Rome, from where we extrapolated only those companies which indicated Rome as their time_zone and which corresponded to the aforementioned basic criteria.

The second consists of the Forbes top 100 Italian companies in terms of profits, market value and assets: http://www.economywatch.com/companies/ forbes-list/italy.html from which we extrapolated only those companies which correspond to the basic criteria listed above.
Furthermore, we did not take into account companies with less than 10,000 followers, in order to limit the number of companies and hence the number of queries to send to the Twitter database. Both sources were combined. Only **20** companies were selected as a result.

The sample has the following characteristics:

Only companies with a minimum of 10,000 followers were taken into account
A maximum of 10,000 followers per company were analysed. I could have selected a proportionately smaller sample according to the number of followers a company has, however I preferred to always query 10,000 users given that the software allowed me to do so.
The sample selection was extracted by the software using a random algorithm which allows true randomness across the entire range of users.

"protected" accounts were not analysed – in other words, those which do not publically share their activities and which therefore cannot be analysed by the software.

Algorithm and method

The algorithm takes into account two classes of parameters. The first corresponds to behaviours which represent characteristics that are probably "human"; the second comprises characteristics which are most likely from "bots".

Characteristics associated with "human" behaviour worth one point:

· The profile contains a name
· The profile contains an image
· The profile contains a physical address
· The profile contains a biography
· The user has at least 30 followers
· The user has been added to a list by other users
· The user has written more than 50 posts
· The user has been geolocalised
· The profile contains a URL
· The user has been included in another user's favourites
· The user uses punctuation in posts
· The user has used a hashtag in their posts at least once
· The user has used an iPhone to log in to Twitter
· The user has used Android to log in to Twitter
· The user has posted with Foursquare
· The user has posted with Instagram
· The user has used the Twitter.com website
· The user has written the userID of another user inside at least one post
· The user has a number of followers which, if doubled, is greater than the number they are following.
· The user publishes content which does not just contain URLs

Characteristics associated with "human" behaviour worth two points:

· At least one post has been retweeted by other users

Characteristics associated with "**human**" behaviour worth **three points**:

· The user has logged into Twitter through different clients

Characteristics associated with "**bot"** behaviour worth **one point**:

· For each characteristic on the "human" list which has not scored points, one "bot" point will be assigned, with the exception of the following:

- o the user has logged in through different clients
- o the user uses the website
- o the user has used Android
- o the user has used iPhone
- o the user has posted with Foursquare
- o the user has posted with Instagram

· User uses only APIs

If any one characteristic of "human" behaviour is true, the corresponding "human" points will be assigned. If it is false, the corresponding "bot" points are assigned.
Conversely, for each "bot" behaviour characteristic, if it is true, "bot" points will be assigned. If it is false, "human" points are assigned.

The software analyses the followers of each brand considered in the sample, and assigns the "human" or "bot" score to each individual follower.
There are some users who present an "uncertain" behaviour, i.e. their behaviour does not feature enough characteristics to identify them as either "human" or "bot" behaviours. The values of these users were not taken into account for the purposes of the summary tables below; however, they were measured and can be seen in the annexes.
Furthermore, there is a marginal number of users who can be defined as "protected", whose activity on this social network cannot be analysed as their privacy settings on Twitter do not allow it. Once again, as above, the values of these users were not taken into account for the purposes of the human/bot assessment, but they were measured and can be viewed in both the tables and the annexes.

Results

The criteria used for subdividing the results defines them as :

- "human" if the "human" value is greater than the "bot" value [Human_Value>bot_Value]
- "uncertain" if the "human" value is lower than the "bot" value with a maximum difference of 4 points.
[Human_Value>bot_Value – 4AND Human_Value<=bot_Value]
- "bot" if the "human" value is lower than the "bot" value by at least 4 points
[Human_Value<=bot_Value – 4]

Table 1 shows a projection of the figures calculated on the total number of followers. Based on the percentage of "bot" behaviours identified in the sample being examined (10,000 followers), the number of followers with "bot" behaviours was projected.
The projection is divided into tables 1a, 1b, 1c according to the various categories analysed by the research.
The tables were ordered based on the percentage of users who have bot type behaviours.

- **Account ID** the username of the brand's account
- **Followers** the number of users which follow the brand
- **Number of bots** number of users which have bot-type behaviours
- **%bot** percentage of users which have bot-type behaviours
- **Number of humans** number of users which have human-type behaviours
- **%humans** percentage of users which have human-type behaviours
- **%protected** percentage of users whose activities cannot be analysed by the software
- **%uncertain** percentage of users which have "uncertain" behaviours

Table 1a Projection on the Total: International companies in the world

Account ID	Followers	Number of bots	%bots	Number of Humans	%Humans	% Protected	%Uncertain
DellOutlet	1520302	699187	**45.99**	460499	30.29	10.52	13.20
WholeFoods	2565369	1137228	**44.33**	1103365	43.01	3.66	9.00
JetBlue	1674437	613514	**36.64**	667765	39.88	11.24	12.24
EA	927690	275524	**29.70**	490562	52.88	6.10	11.32
YSL	904271	222903	**24.65**	494817	54.72	8.37	12.26
SouthwestAir	1297848	300841	**23.18**	697723	53.76	12.10	10.96
threadless	1847061	350572	**18.98**	1095122	59.29	10.49	11.24
pepsi	755547	119905	**15.87**	458164	60.64	11.30	12.19
BlackBerry	977437	150037	**15.35**	604545	61.85	11.07	11.73
CocaCola	548512	72020	**13.13**	320605	58.45	11.94	16.48
PlayStation	1297768	150671	**11.61**	831869	64.10	12.99	11.30
SamsungMobile	1286753	136010	**10.57**	790967	61.47	14.98	12.98
Starbucks	2529115	174005	**6.88**	1695545	67.04	16.89	9.19

Table 1b Projection on the Total: International companies in Italy							
Account ID	Followers	Number of bots	%bots	Number of Humans	%Humans	% Protected	%Uncertain
IKEAITALIA	119736	54983	**45.92**	35825	29.92	9.54	14.62
VodafoneIT	155050	60113	**38.77**	58981	38.04	9.86	13.33
3Italia	78266	28019	**35.80**	32825	41.94	8.98	13.28
nokiaitalia	105708	37738	**35.70**	42875	40.56	11.05	12.69
CiscoIT	19745	5426	**27.48**	7936	40.19	14.35	17.98
fnac_italia	12818	1630	**12.72**	8378	65.36	12.00	9.92

Table 1c Projection on the Total: Italian companies							
Account ID	Followers	Number of bots	%bots	Number of Humans	%Humans	% Protected	%Uncertain
Treccani	84921	37934	**44.67**	29485	34.72	8.58	12.03
librimondadori	153045	65442	**42.76**	53857	35.19	8.80	13.25
coinstore	35630	15182	**42.61**	12128	34.04	9.56	13.79
LaFeltrinelli	294223	124839	**42.43**	101272	34.42	9.21	13.94
feltrinellied	59796	23243	**38.87**	23261	38.90	9.44	12.79
dolcegabbana	347462	97602	**28.09**	161257	46.41	12.63	12.87
telecomitaliaTw	56402	15133	**26.83**	27710	49.13	11.59	12.45
fiatontheweb	18466	3296	**17.85**	11888	64.38	6.43	11.34
Alitalia	17505	2974	**16.99**	9822	56.11	13.27	13.63
lamborghini	39347	6496	**16.51**	22605	57.45	13.26	12.78
Tim_Official	227331	33827	**14.88**	127851	56.24	13.05	15.83
pirelli_media	24505	3614	**14.75**	17220	70.27	4.77	10.21
Armani	68674	9848	**14.34**	39460	57.46	14.56	13.64
Yoox	13903	1809	**13.01**	10096	72.62	5.92	8.45
MaisonValentino	23464	2680	**11.42**	14806	63.10	15.46	10.02
LuisaViaRoma	24030	2408	**10.02**	17184	71.51	9.55	8.92
Ferragamo	24440	2371	**9.70**	16355	66.92	14.20	9.18
EmilioPucci	31961	2761	**8.64**	21427	67.04	15.26	9.06
GFISoftware	38185	1848	**4.84**	32583	85.33	1.58	8.25
PomodoroMutti	15501	370	**2.39**	14058	90.69	1.99	4.93

Tables 2a, 2b and 2c represent the data calculated by the software on the sample of 10,000 users per brand, containing eight columns ordered by percentage of "bots":

- **Account ID** the username of the brand's account
- **Users considered** the number of users chosen randomly by the software
- **Number bots** number of users which have bot-type behaviours
- **%bot** percentage of users which have bot-type behaviours
- **Number humans** number of users which have human-type behaviours
- **%humans** percentage of users which have human-type behaviours
- **%protected** percentage of users whose activities cannot be analysed by the software
- **%uncertain** percentage of users which have "uncertain" behaviours

Table 2a: Data on the sample International companies in the world							
Account ID	Users Considered	Number bot	%bot	Number humans	%humans	%protected	%uncertain
DellOutlet	10000	4599	**45.99**	3029	30.29	10.52	13.20
WholeFoods	10000	4433	**44.33**	4301	43.01	3.66	9.00
JetBlue	10000	3664	**36.64**	3988	39.88	11.24	12.24
EA	10000	2970	**29.70**	5288	52.88	6.10	11.32

YSL	10000	2465	**24.65**	5472	54.72	8.37	12.26
SouthwestAir	10000	2318	**23.18**	5376	53.76	12.10	10.96
threadless	10000	1898	**18.98**	5929	59.29	10.49	11.24
pepsi	10000	1587	**15.87**	6064	60.64	11.30	12.19
BlackBerry	10000	1535	**15.35**	6185	61.85	11.07	11.73
CocaCola	10000	1313	**13.13**	5845	58.45	11.94	16.48
PlayStation	10000	1161	**11.61**	6410	64.10	12.99	11.30
SamsungMobile	10000	1057	**10.57**	6147	61.47	14.98	12.98
Starbucks	10000	688	**6.88**	6704	67.04	16.89	9.19

Table 2b: Data on sample International companies in Italy

Account ID	Users Considered	Number bot	%bot	Number humans	%humans	%protected	%uncertain
IKEAITALIA	10000	4592	**45.92**	2992	29.92	9.54	14.62
VodafoneIT	10000	3877	**38.77**	3804	38.04	9.86	13.33
3Italia	10000	3580	**35.80**	4194	41.94	8.98	12.69
nokiaitalia	10000	3570	**35.70**	4056	40.56	11.05	13.28
CiscoIT	10000	2748	**27.48**	4019	40.19	14.35	17.98
fnac_italia	10000	1272	**12.72**	6536	65.36	12.00	9.92

Table 2c: Data on sample Italian companies

Account ID	Users Considered	Number bot	%bot	Number humans	%humans	%protected	%uncertain
Treccani	10000	4467	**44.67**	3472	34.72	8.58	12.03
librimondadori	10000	4276	**42.76**	3519	35.19	8.80	13.25
coinstore	10000	4261	**42.61**	3404	34.04	9.56	13.79
LaFeltrinelli	10000	4243	**42.43**	3442	34.42	9.21	13.94
feltrinellied	10000	3887	**38.87**	3890	38.90	9.44	12.79
dolcegabbana	10000	2809	**28.09**	4641	46.41	12.63	12.87
telecomitaliaTw	10000	2683	**26.83**	4913	49.13	11.59	12.45
fiatontheweb	10000	1785	**17.85**	6438	64.38	6.43	11.34
Alitalia	10000	1699	**16.99**	5611	56.11	13.27	13.63
lamborghini	10000	1651	**16.51**	5745	57.45	13.26	12.78
Tim_Official	10000	1488	**14.88**	5624	56.24	13.05	15.83
pirelli_media	10000	1475	**14.75**	7027	70.27	4.77	10.21
Armani	10000	1434	**14.34**	5746	57.46	14.56	13.64
Yoox	10000	1301	**13.01**	7262	72.62	5.92	8.45
MaisonValentino	10000	1142	**11.42**	6310	63.10	15.46	10.02
LuisaViaRoma	10000	1002	**10.02**	7151	71.51	9.55	8.92
Ferragamo	10000	970	**9.70**	6692	66.92	14.20	9.18
EmilioPucci	10000	864	**8.64**	6704	67.04	15.26	9.06
GFISoftware	10000	484	**4.84**	8533	85.33	1.58	8.25
PomodoroMutti	10000	239	**2.39**	9069	90.69	1.99	4.93

The detailed list of the data analysed, with the usernames concealed for the purposes of confidentiality, is available at the following address :
http://iulm.camisanicalzolari.com/MCCricercaUserlistSource.zip
so that it can be used by other researchers wishing to apply any new methods or algorithms they may have. The compressed folder contains the various files subdivided by category.

Each of these files contains, for each brand and for each individual follower, the following headings:
> Father_ScreenName: The Twitter name of the Brand account
> Father_Following: The number of Twitter accounts that the Brand if following
> Father_Followers: The number of followers of the Brand
> Follower_Id: the Twitter_id of the user in the Panel
> Follower_ScreenName: The Twitter name of the user in the Panel
> Follower_Protected: If TRUE the user in the Panel is private and has no statistics
> Follower_HumanValue: The Human Points of the user in the Panel
> Follower_BotValue: The Bot Points of the user in the Panel

Conclusions

A very high number of users with "bot" behaviours was found in certain companies, with percentages in excess of 45%, despite the fact that, as described previously, the algorithm allowing "human" and "bot" points to be assigned was defined with very conservative parameters.
Regardless of the importance I decided to place on the individual parameters used, it is interesting to observe that the algorithm works well, as it produced very different results depending on which brand it was applied to.
In the same table, for numerical samples of similar sizes we find DellOutlet with 45.99% "bots", and Starbucks with 6.88%.
We can deduce that users display profoundly different behaviours, revealing large numbers of "inactive" users with behaviours which, if we consider the algorithm and its weighted parameters as being valid, can reasonably be considered non-human behaviours.