

Identifying Instagram Crowdturfers with Machine Learning

Originally published at <https://www.unite.ai/identifying-instagram-crowdturfers-with-machine-learning/>



Researchers in Italy and Iran claim to have formulated the first machine learning system capable of recognizing the 'crowdturfing' activity of human (rather than automated) influencer accounts on the Instagram platform. Crowdturfers are real people who perform 'profile building' services to platforms which sell such activity on a wholesale basis.

The new method claims an accuracy score of around 95%, and uses semi-supervised learning in Natural Language Processing (NLP) systems.

The authors claim that to the best of their knowledge, their system represents the first crowdturfing (CT) detector system that can reliably hone in on non-bot accounts that are engaged in fake, paid profile engagement and boosting.

To accomplish this, the authors purchased 1293 crowdturfing profiles from 11 CT platform providers in order to obtain data to train their CT detector. Since Instagram has a number of effective anti-bot measures in place, the researchers note, those seeking to exploit the platform's enormous user base for commercial purposes have turned to paying genuinely influential Instagrammers to 'engage strategically' with 'client' accounts, mostly by sharing comments, or through activity related to comments on posts.

Having trained the model, the authors then set it loose to analyze the engagement profiles of 20 'mega-influencers', each with over 1 million followers, concluding that *'more than 20% of their engagement was artificial'*.

The [paper](#) is titled *Are We All in a Truman Show? Spotting Instagram Crowdturfing through Self-Training*, and comes from five researchers across the University of Padova in Italy, and Iran's Imam Reza University.

Breaching the Instagram TOS

Unlike Twitter, favored by social media researchers due to its commitment to aiding research, Instagram not only provides no API or updated data dumps to help researchers, but prohibits machine-driven browsing in its Terms of Service. Therefore the researchers' first task was to gain an exemption from their guiding Institutional Review Board, justified by [prior works](#) that used a similar approach to investigate 'underground activities'.

The crowdturfing services were purchased for fresh Instagram accounts created by the researchers for their purposes, all of which were deleted after the experiment, obviating the involvement of 'legitimate' users. Neither the influencer accounts studied nor the CT platform services are named.

Another ethical hurdle was that the researchers could not request consent of the influencers being studied, due to the [Hawthorne effect](#) (i.e. it might have changed the influencers' behavior), and this exemption was also granted by the IRB.

Finally, since Instagram allows 'manual collection' of data, the researchers compromised on their breach of the TOS by setting their automated scraping tools to 'human speed', which necessitated a data-gathering phase of five months.

Humans for Sale

The researchers purchased 100 'fake follower' profiles from each of 11 (unnamed) providers.

The paper states*:

'All the providers we selected ensure to deliver followers who interact with the target profiles by liking and commenting on their posts to boost their engagement rate.'

'These CT profiles are identified as high quality followers and usually cost more than "base" fake profiles. The reliability of these providers is supported by famous [review] platforms like TrustPilot.'

Provider	Price	Delivery Time	Drop Protection	Followers Received	Followers 1 Month	#Followers Avg (std)	#Following Avg (std)	Private Profiles	#Posts Avg (std)	URLs in Biography
CT-1	\$5.69	Instant	yes	115	74	409.59 (1110.46)	812.38 (1331.52)	0.13%	14.83 (57.98)	0.08%
CT-2	\$2.39	5-10m	no	211	340	44.61 (108.85)	4679.75 (1452.19)	0%	16.0 (18.06)	0%
CT-3	\$2.95	Instant	yes	111	85	132.17 (327.28)	3027.08 (1883.18)	0.05%	20.19 (55.99)	0.09%
CT-4	\$2	Instant	no	100	42	239.45 (262.64)	2735.6 (1286.65)	0.45%	111.95 (332.2)	0.01%
CT-5	\$3.95	Gradual	yes	79	61	201.43 (214.0)	3510.77 (2316.12)	0%	16.06 (12.13)	0.054%
CT-6	\$2.89	24-72h	yes	136	129	36.79 (39.64)	2398.88 (2191.18)	0%	14.06 (5.69)	0%
CT-7	\$2.70	1h	yes	108	109	39.23 (73.32)	3966.36 (761.16)	0%	19.74 (20.13)	0%
CT-8	\$5.78	Gradual	no	110	95	57.52 (138.97)	1818.84 (1353.95)	0.04%	29.75 (41.09)	0.01%
CT-9	\$3.95	12h	no	109	99	129.54 (759.85)	2012.93 (1198.17)	0.06%	26.99 (74.94)	0%
CT-10	\$5.94	Gradual	no	97	94	83.38 (174.57)	2118.31 (1323.78)	0.03%	40.28 (51.5)	0%
Low quality	\$0.80	24-72h	no	117	96	87.26 (276.26)	3200.67 (3041.89)	0.04%	1.88 (6.15)	0.02%
Real	-	-	-	-	-	359.33 (237.87)	571.24 (517.53)	57.92%	279.09 (369.67)	14.44%

From the paper, statistics on the (anonymized) CT platform providers, each a marketplace for 'corrupted' real-world influencer accounts. This table outlines information retrieved by the researchers through the analysis of the 100 profiles purchased from each source. Source: <https://arxiv.org/pdf/2206.12904.pdf>

The average cost of buying an Instagram influencer, the paper notes, is not that high, at approximately \$3 for 100 'high quality' followers. The authors note:

'Most providers deliver the followers within a few hours. They offer a drop protection, which means that the number of followers the customer purchases will either remain stable over time or new followers will be delivered to replenish the lost ones.'

The researchers report that some of their fresh Instagram accounts suffered a loss of 15-20% of CT followers after one month, but that in certain cases they gained more than expected. For the most expensive CT provider (CT-10, in the table above), only three followers were lost after one month.

The paper notes that the followed/following ratio becomes more 'authentic' the more you pay to the CT provider, with the second-most expensive provider offering a ratio that's very close to a standard user's baseline.

One characteristic of a CT Instagram account is that its profile will rarely be set to 'private' (a fact that enabled data to be drawn from the purchased fake followers, since most of the analyses centered on profiles and related comments), though this should not be seen as a reliable 'signal' in this regard.

'People joining these platforms are interested in generating a minimum amount of posts that make them reliable, except few cases (CT-4, CT-10). The low-quality profiles show a very high imbalance in followers and following, and the average number of posts is close to 0, far below the CT profiles.'

Data

The researchers collected data through an implementation of the browser-automating framework Selenium. The resulting dataset includes profile information from 1293 CT and 1307 non-CT users.

This admittedly low sample quantity made it feasible to set Selenium to a credibly human speed over a rational period of time. Additionally, the authors note, the representative/interpretive power of semi-supervised learning techniques accommodates smaller datasets very well. Having experimented, for the purposes of thoroughness, with a fully-supervised model, the researchers conclude:

'[The] results in the semi-supervised mode do not differ significantly from those in a supervised way. This suggests that CT profiles share very similar [characteristics], and that the algorithm can converge [through a small amount of] labeled data.'

The authors gathered all available data from the source code of the 'compromised' users' profile pages, including details generally obscured when rendered, such as the #videos element.

They then pre-processed the data features by removing those with zero or low variance, and finally converted any categorical or non-numeric data into strictly numeric or Boolean features.

Numeric Features	Boolean Features
# followers	Account is private
# following	Account is verified
# videos	Account has clips
# posts	Account is business account
# characters in username	Account has external URLs
# characters in fullname	Accounts has category name
# characters in biography	Account has multiple categories
# digits in username	
# non-alphabetic char in fullname	
# hashtags and mentions in biography	

Characteristics of the final dataset.

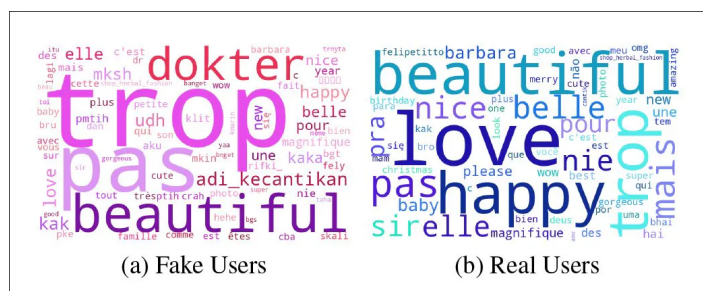
Method and Explorations

Besides, Selenium, technologies used across the experiments include: a version of SpaCy implemented with a transformer-based pipeline; a scikit learn [self-training classifier](#); and the [Instaloader](#) framework.

There is no customary 'results' section in the new paper, since it deals with an objective (i.e., automated inference of corrupt Instagram accounts) that veers away from the central locus of interest to date (i.e., automated inference of automated bot activity on Instagram), meaning that there is no like-for-like prior work against which to compare it.

The researchers adopted a wide range of methods on the available purchased users, (which they feel comfortable describing as 'fake' rather than just 'non-CT', since these genuine accounts are conducting non-organic, paid engagement activities), across a range of NLP-related technologies.

Among the facets studied were language analysis (which, in the CT world, nearly always defaults to English, though CT platforms offer geo-located non-English followers too); comment counts (where fake users stick very close to the frequency of real users, for fear of detection); and common words analysis:



Word clouds from fake and real users

The paper notes that the prevalence of the word 'dokter' (see image above) in fake accounts seems to relate to a specific internal campaign:

“Dokter” [appeared] in 1069 distinct comments. By further investigating the accounts spamming [this] word, we found a small portion of what seems to be a botnet whose objective is to spam “Instagram doctors” accounts. All these doctors’ profiles have a WhatsApp business link that, once clicked, starts a chat with a message to complete.’

As far as the researchers can deduce, this strange artifact may be a remnant of a large botnet that they stumbled across while seeking activities from real Instagram users.

In total the researchers collected 603,007 comments from posts across 248,388 unique Instagram users, of which, the authors estimate, 55,719 were crowdturfing accounts.

The paper notes with interest the dominance of female-themed topics in the gathered data. Having used [GPU-PDMM](#) (a technique developed for the obligatory short posts on Twitter) to extract 12,830 suitable comments from an available corpus of 121,822 comments, the algorithm found that in considering content from 12 males and 8 females, the majority of comments deal with female-related topics.

<i>Comments</i>	<i>Top words for topic</i>	<i>Label</i>
2526	beautiful queen cute pretty sexy girl hot gorgeous baby cool sweet angel	Female Sexy
2490	wow gorgeous beauty stunning amazing super perfect awesome goddess angel top	Female Goddess
1879	nice beautiful very good looking picture pic love lovely sweet smile photo	Photo
1744	you love look beautiful like much god all baby with too always sir both pretty	Both genders Compliments
1090	happy new year birthday diwali halloween thanksgiving christmas republic belated	Holidays
699	best one eyes most world woman ever amazing look her what dear	Female Beautiful
613	your see life more enjoy liked hope every moment smiles	Life
607	like big sir fan i'm bro looks yes have work real	Male Fitness
604	you and get well man can great soon wish day video better keep	Male Recovery
578	kiss please red with lips pretty face eyes bite follow lip-lock	Female Face

The top 10 topics extracted from fake topics in one of the researchers' experiments.

The researchers conclude:

'[While] Instagram and the research community focused a lot on detecting bots and automated accounts, we believe more studies should be conducted on CT activities, which negatively impact influencer marketing, the Instagram platform, and most of its users.'

* Researchers' quoted TrustPilot URL omitted.

First published 28th June 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.

Portfolio site: martinanderson.ai

Contact: martin@martinanderson.ai



Discover More