Are We All in a Truman Show? Spotting Instagram Crowdturfing through Self-Training

Pier Paolo Tricomi*‡, Sousan Tarahomi†, Christian Cattai*, Francesco Martini*, Mauro Conti*‡

*Department of Mathematics, University of Padua, Padua, Italy

{ tricomi,conti}@math.unipd.it {christian.cattai,francesco.martini.11}@studenti.unipd.it

†Department of Computer Science, University of Twente, Enschede, Netherlands – s.tarahomi@utwente.nl

‡Chisito S.r.l., Padua, Italy

Abstract—Influencer Marketing generated $16 billion in 2022. Usually, the more popular influencers are paid more for their collaborations. Thus, many services were created to boost profiles' popularity metrics through bots or fake accounts. However, real people recently started participating in such boosting activities using their real accounts for monetary rewards, generating ungenuine content that is extremely difficult to detect. To date, no works have attempted to detect this new phenomenon, known as crowdturfing (CT), on Instagram.

In this work, we propose the first Instagram CT engagement detector. Our algorithm leverages profiles' characteristics through semi-supervised learning to spot accounts involved in CT activities. Compared to the supervised approaches used so far to identify fake accounts, semi-supervised models can exploit huge quantities of unlabeled data to increase performance. We purchased and studied 1293 CT profiles from 11 providers to build our self-training classifier, which reached 95% F1-score. We tested our model in the wild by detecting and analyzing CT engagement from 20 mega-influencers (i.e., with more than one million followers), and discovered that more than 20% was artificial. We analyzed the CT profiles and comments, showing that it is difficult to detect these activities based solely on their generated content.

Index Terms—Crowdturfing Detection, Fake Engagement, Instagram, Fake Profiles, Collusion, Self-Training

I. INTRODUCTION

Instagram (IG) is the most popular photo-sharing social media, with around 1.5 billion monthly active users [45], and the preferred platform for influencer marketing [23]. Unfortunately, such a market is often manipulated, making influencers unreliable [34]. Indeed, many providers offer services to boost the visibility and fame of a specific account, for example, by increasing the number of followers, likes, and comments. As (social) bots [12] or fake accounts [41] originally conducted these activities, IG has adopted Machine Learning (ML) algorithms to remove them efficiently. Instead, nowadays, *real people use their accounts to conduct such unauthentic activities behind a monetary reward*. In the literature, this collusive phenomenon is called crowdturfing (CT), a term combining the collaboration of many individuals (crowdsourcing [1]) with an apparently natural action controlled by agencies (astroturfing [20], [51]). Figure 1 shows Fake and CT profiles. While the fake profile exhibits well-known characteristics (e.g., no posts, no bio, few followers [41]), the CT profile looks legit (indeed, it is a real-person account), and thus more challenging to spot. Considering CT engagement is not real, we can label it as fake.

Fake engagement damages the authenticity of social media, creating threats such as brand abuse or followers farming [54].
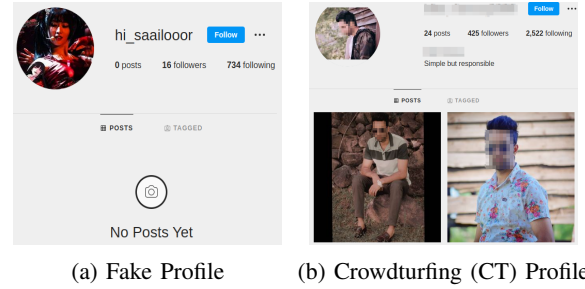

(a) Fake Profile   (b) Crowdturfing (CT) Profile

Fig. 1: Example of fake vs crowdturfing profiles.

To spot fake engagement derived from CT, a reliable strategy is to detect the involved accounts. Different approaches have been proposed to distinguish between genuine or fake users, but to the best of our knowledge, none consider CT-involved accounts on IG. Among these approaches, machine learning-based solutions are the most powerful and cost-effective techniques [38]. While most proposed ML algorithms for account classification leverage supervised learning, there is always the need for an adequately labeled ground truth, which is inherently difficult to obtain for CT activities [44]. Instead, Semi-Supervised Learning (SSL) methods could be more appropriate when only a few labeled samples are available or needed. In fact, a large amount of unlabeled data can help improve the classification without impacting the performance [43]. Last, given the intrinsic differences between IG and other social media where CT has been studied (e.g., Twitter), we run our detector in the wild to analyze CT profiles and their fake engagement under several aspects, highlighting the difficulty of detecting such activities merely by looking at generated content.

Contribution. Our contribution is summarized as follows:

- We are the first, to the best of our knowledge, to propose a CT engagement detector on IG, which furtherly reduces the computational costs of previous fake accounts and bot detectors by leveraging semi-supervised algorithms;
- We provide a detailed analysis of CT providers to explore the services they offer and the profiles involved;
- We analyze Instagram CT engagement in the wild, mainly

related to comments, by running our detector on 1000 posts generated by 20 mega-influencers.

- Our (anonymized) data will be released upon request to help researchers study CT activities on IG.

Structure. §II presents related works. §III examines CT providers, while §IV describes our CT detection mechanism. CT profiles and comments spotted in the wild are analyzed in §V and §VI, respectively. §VII concludes the paper.

II. RELATED WORKS

We discuss CT detection in social media(§II-A), along with fake account detection on IG (§II-B) and the adoption of semi-supervised algorithms to detect bots and fake profiles (§II-C).

A. Crowdturfing in Online Social Media

Researchers have examined CrowdTurfing (or *collusive* [54]) social media activities first on Sina Weibo, and then primarily on Twitter, on which misinformation or political campaign manipulation often occurs. Wang et al. [51] investigated two popular crowd-sourcing sites in china and tracked down the CT campaigns on Sina Weibo. Then, they discussed the characteristics of CT and genuine accounts and analyzed the CT campaigns. Another work on Sina Weibo [53] examined CT accounts engaging in political activities, claiming their methodology could not find any clear evidence to show the presence of large-scale political CT. The authors of [32] categorized different types of CT tasks on Fiverr and applied ML algorithms to distinguish these tasks from legitimate ones. Song et al. [44] focused on spotting targets of CT tasks, such as pots, pages, and URLs, on Twitter. Chetan et al. [9] proposed CoReRank, an unsupervised method for detecting suspicious tweets and collusive retweeters. Dutta et al. developed several mechanisms to detect and characterize collusive users involved in black market services on Twitter [14]–[16]. Eventually, the authors in [50] qualitatively investigated the impact of CT activity on content visibility and popularity on IG. They claimed that IG is vulnerable to CT activities and stressed the need for a CT detector. To the best of our knowledge, we are the first to implement such a detector for IG, adopting a performing and efficient SSL approach.

As outlined in [17], social media have unique characteristics, purposes, and interactions that require tailored CT studies. For instance, researchers have recently moved their interests to YouTube [18], [19], showing that platforms besides Twitter need to be studied. We argue that IG is fundamentally different from Twitter. First, IG has roughly 1.5 Billion monthly active users (three times Twitter ones [45]), who spend three times the time spent on Twitter [52], indicating IG's greater influence (2022). Second, they serve very different purposes [36]: Twitter lets users communicate in an elevator pitch fashion with quick messages, while IG primarily focuses on creating interactive communities through images and videos. Not surprisingly, nearly 80% of brands use IG influencers for their marketing campaigns, compared to 20% on Twitter [23], [27]. Last, while Twitter APIs[1] allow collecting a variety of

users' data (e.g., profile info, activities, connections), IG APIs[2] release only limited data. Due to these reasons, algorithms developed on Twitter cannot inherently apply to IG, so deploying methods to detect IG CT activities is urgently needed.

B. Instagram Fake Accounts Detection

Although no prior works attempted to detect CT activities on IG, several works tried detecting fake profiles [41] or (social) bots [12], [47], which we can refer to as *classic* fake profiles. In [48], the authors developed an ML model to detect fake likes on IG, deploying honeypots and botnets to collect the ground truth. They employed ML methods to find the authenticity of likers with features including the number of followers, following, and their relationships. To detect fake and automated IG accounts, the authors in [2] applied different ML algorithms on posts and media-related features, obtaining 86% and 94% accuracy for automated and fake accounts, respectively. In [42], the authors used bagged decision trees on profile-related features to detect trivial (manually labeled) fake users. Zarei et al. [54] applied clustering methods to track down impersonators in three different categories based on their profile similarity. In [39], the authors tried to detect three categories of fake accounts: active, inactive, and spammers. They bought fake accounts from Indonesian providers; however, most were simple bots, not linked to CT activities. They reached 92% accuracy using Random Forest. Kim and Hany [30] proposed a neural network to detect engagement bots by three sets of features, including text, behavior, and graph-based features. Given the existence of fake accounts and bots detection mechanisms in the literature, we will evaluate such methods on CT profiles, understanding to which extent *classic* fake accounts differ from CT accounts.

C. Semi-supervised Fake Accounts Detection

SSL approaches can leverage a vast amount of unlabeled data, reducing labeling costs with few to no drops in performances [43]. Most of these approaches were adopted on social media to detect Sybil Nodes or Bots. SybilBelief [24] is an SSL framework for finding Sybil nodes such as spammers and impersonators. SybilTrap [3] uses label propagation random walk as a semi-supervised transductive-learning approach to detect malicious users. This approach focuses on both structural and content-based features. Dorri et al. [13] developed SocialBotHunter as an SSL collective classification technique to detect social bots in Twitter-like platforms. Their approach uses the social behavior and interaction of users. Last, SEMIPSM [5] is an SSL Laplacian SVM model using manifold regularization to discover users responsible for propagating misinformation on social media.

III. CROWDTURFING PROVIDERS ANALYSIS

To spot CT activities, such as fake followings or comments, we must study, understand, and collect "authentic" CT profiles. Previous studies on fake profiles detection collected fake

---

[1] https://developer.twitter.com/en/docs/twitter-api

[2] https://developers.facebook.com/docs/instagram-api/

TABLE I: Characteristics of Crowdturfing providers. The table reports information claimed by the provider and retrieved by analyzing 100 profiles bought from each. The last row reports info on real profiles for comparison.

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 (106.85)	4679.75 (1452.19)	0%	16.0 (8.06)	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%

profiles or bots by manually searching for poorly designed accounts, such as those without a profile pic, with alpha-numeric names, or a very low number of posts and followers [2], [42]. Other works focused on synthetic data [48] or bought mostly naive fake accounts from local providers [39]. However, the profiles gathered using these methodologies indubitably introduce bias in the data, and the resulting detectors will identify just simple profiles, very likely driven by a bot master.

Instead, we are interested in spotting fake activities conducted by real people profiles that are populating and remaining on IG by evading its bot detection mechanisms [28], [29], [37]. To this aim, we selected 10 well-known crowdturfing providers and bought from each of them 100 fake followers. All the selected providers **ensure to deliver real followers (i.e., real people)** 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 (i.e., profiles usually managed by a bot master). To identify reliable providers, we selected services that had at least an average of 3 (out of 5) stars on the famous reviews platforms TrustPilot[3]. Moreover, many of our CT providers allow people to freely join their platforms to participate in CT activities, confirming the reliability of the service and the presence of human activity behind the fake engagement they generate. Table I describes these providers, along with information about a low-quality provider. We also included information on real profiles we used in our study.[4] To limit CT activities on IG, we bought CT followers for profiles we created for the study, which we deleted at the end. We are not reporting the names of the CT providers to avoid the encouragement of such activities.

The table shows that the price average is pretty low, around $3 for 100 high-quality followers, but much higher than the $0.80 for 100 low-quality followers. Followers are usually delivered within a few hours, and most providers offer drop protection, replenishing any lost follower. To assess the providers' reliability, we checked how many followers remained after one month. On average, we lost only 15-20% of them, and sometimes, we gained more. CT-4, the

least expensive provider, lost the most, while we lost only 3 followers from the most expensive CT-10. Compared to real profiles, CT profiles have a noticeable difference in followers and following. This is understandable, given that the more they follow and interact, the more they get paid. However, from CT-1, the second most expensive provider, the follower/following balance is quite close to real profiles. The CT profiles also are quite different from real ones in terms of being private, the number of posts, and the URLs in the biography. Very likely, CT platforms require profiles to be public. People joining these platforms generate a minimum amount of posts to be 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. Among the properties we did not report in the table, some providers allow customers to increase their followers periodically or buy followers from a specific geographical region or language.

IV. CROWDTURFING PROFILES DETECTION

Instead of directly detecting CT activities (e.g., a fake comment), we first detect profiles involved in CT activities, and accordingly, we label their interaction as CT. The rationale behind this approach is that CT profiles are mainly **real accounts** belonging to individuals willing to create fake interactions. Thus, their interactions should resemble genuine ones, in both content and temporal activity [17]. Similarly, their profile information should appear legitimate, which makes detecting CT profiles considerably different from spotting *classic* fake profiles [14] (i.e., the focus of previous works). Indeed, the latter usually present simplistic features (e.g., no posts, no followers), or recognizable patterns (e.g., low-variability content) [41]. We now present the dataset we collected to classify CT profiles (§IV-A), our detection model (§IV-B), and a comparison with previous approaches (§IV-C).

A. Dataset and feature selection

Since there are no IG CT datasets available, we collected our own. Given IG API could not provide our requirements, we performed automated data collection through Selenium[5]. For our detector, we use general profile info (e.g., #followers,

---

[3]https://www.trustpilot.com/

[4]Here, to simplify comparisons, we excluded celebrities and highly-followed accounts (> 500 followers), which could present inflated statistics.

[5]https://www.selenium.dev/

#following, #posts) instead of behavioral patterns since IG does not provide such information, unlike other social media (e.g., Twitter). Some previous works [39], [48] used features that are not publicly available, e.g., the number of likes of a user's posts, limiting their approach only to public profiles. Instead, we focused only on profile features that are publicly available for both public and private profiles.[6]

The dataset contains the profile information of 2600 users, including 1293 CT and 1307 authentic accounts. The CT profiles are the ones analyzed in Section III. We gathered authentic accounts similarly to previous works [2], [39], [42]. We included from several countries and fields: general users from our expanded social connections, verified or business accounts, and celebrities. Three authors validated these accounts through extensive manual labeling, adopting a majority voting for the decision, and focusing on attributes such as the Follower/Following imbalance, the number of posts, or the full name. The feature distributions of our real accounts (Table I) closely align with previous works. For the collected accounts, we gathered all the attributes available on the profile page. Then, we pre-processed the features by removing those with zero or very low variance. Last, we transformed categorical and non-numeric attributes into numeric or boolean features. The final features are shown in Table II. Since all the data we collected is public, we will make it available upon request (anonymized) to help the research community studying CT.

TABLE II: Final set of features of our dataset.

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

B. Our Semi-Supervised Model

The next step is to develop a detector to distinguish between real and CT profiles. In light of previous discussions, labeling CT profiles is challenging [44]. Therefore, instead of adopting supervised methods as in previous works, we use SSL to maximize the use of unlabeled data and improve generalization. While most previous SSL approaches on social networks utilized graph-based methods, we consider only profile-related features, making our model less complicated and easier to handle (e.g., for practitioners). Our self-training approach is depicted in Figure 2. Initially, we divide the dataset into labeled and unlabeled datasets, discarding all labels from the unlabeled dataset. In the first training cycle (dashed arrows in the figure), training data corresponds to labeled data. We train a classifier with this data and ask it to predict the labels of all the unlabeled samples, generating their pseudo-labels. For each pair *sample:pseudo-label*, we check the prediction probability (i.e., classifier confidence, from 0 to 1) associated with the pseudo-label. If the probability is higher

than 0.75, we add the pair *sample:pseudo-label* to the training data; otherwise, the sample remains unlabeled. We repeat the training cycle (train the classifier → predict pseudo-labels → enlarge the training set) 10 times or until no unlabeled data remains. The final model corresponds to the classifier of the last iteration.
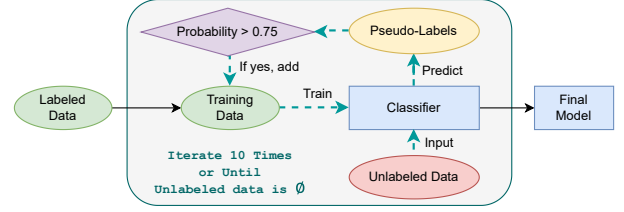

Fig. 2: Schema of Self-Training process. Dashed arrows represent the training cycle.

We implemented our models using Scikit-learn.[7] We randomly split our dataset in a stratified mode to have 80 percent of data for training (2080 labeled samples) and the remaining for testing. We applied 5-StratifiedKFold Cross-Validation on the training set to find the best model hyper-parameters. In each iteration, one fold (∼416 samples) was left out to validate the model, while the remaining ∼1664 samples were used to train the classifier in the semi-supervised fashion described above. To demonstrate the power of SSL, we tested different (small) labeled training data portions: 1, 3, 5, and 9%. As classifiers and hyper-parameters, we tested:

- K-Nearest Neighbor (KNN): *n_neighbors*=[1, 3, 5, 10];
- Logistic Regression (LR): *penalty*=[none, l1, l2], *C*=[10, 1, 0.1], *solver*= [lbfgs, liblinear];
- Decision Tree (DT): *max_depth*=[none, 3, 5, 10], *samples_leaf*=[1, 3, 5, 10];
- Random Forest (RF): *max_depth*=[none, 3, 5, 10], *samples_leaf*=[1, 3, 5, 10], *n_estimators*=[10, 100].

The best hyper-parameters for each model were selected through a grid-search approach. We also trained the classifiers on all the labeled data (i.e., in a supervised mode) for comparison. The results for each classifier using the best hyper-parameters during the cross-validation are reported in Table III.

The table shows that increasing labeled data does not necessarily improve the model's performance but increases its stability. Moreover, the results in the SSL mode do not differ significantly from supervised ones. This suggests that CT profiles share similar characteristics, as partially discussed in §III, and algorithms can converge by taking a few labeled data. On the contrary, adding more samples could lead to overfitting or biasing the classifier, reducing prediction accuracy (as happened for LR 0.03). The LR classifier with 1 percent of labeled data (penalty = *l2*, C = 1, solver = *liblinear*) showed the best cross-validation results among the semi-supervised models, so it was selected as the final model.[8] Such a model reached 95% accuracy and F1-score on the test set and was used in the remainder of our analyses.

---

[6]Some attributes (e.g., #videos) were retrieved from the page source code.

[7]https://scikit-learn.org

[8]We discarded RF sup. (same scores) since the paper focuses on SSL. Practitioners should choose models with the best performance.

TABLE III: Average±std of classification results of the best models during cross validation. Sup = Supervised.

Model and % Labels Used	Train Accuracy	Valid. Accuracy	Valid. Precision	Valid. Recall	Valid. F-Measure
KNN *0.01*	$0.79_{\pm 0.04}$	$0.79_{\pm 0.03}$	$0.84_{\pm 0.02}$	$0.79_{\pm 0.04}$	$0.78_{\pm 0.04}$
0.03	$0.92_{\pm 0.04}$	$0.92_{\pm 0.02}$	$0.92_{\pm 0.03}$	$0.92_{\pm 0.04}$	$0.92_{\pm 0.04}$
0.05	$0.93_{\pm 0.02}$	$0.94_{\pm 0.01}$	$0.93_{\pm 0.02}$	$0.93_{\pm 0.02}$	$0.93_{\pm 0.02}$
0.07	$0.92_{\pm 0.00}$	$0.92_{\pm 0.00}$	$0.92_{\pm 0.00}$	$0.92_{\pm 0.01}$	$0.92_{\pm 0.00}$
0.09	$0.96_{\pm 0.01}$	$0.95_{\pm 0.00}$	$0.96_{\pm 0.01}$	$0.96_{\pm 0.01}$	$0.96_{\pm 0.01}$
Sup.	$0.97_{\pm 0.01}$	$0.97_{\pm 0.00}$	$0.97_{\pm 0.01}$	$0.97_{\pm 0.01}$	$0.97_{\pm 0.01}$
LR *0.01*	$\mathbf{0.97}_{\pm 0.01}$	$\mathbf{0.97}_{\pm 0.00}$	$\mathbf{0.97}_{\pm 0.01}$	$\mathbf{0.97}_{\pm 0.01}$	$\mathbf{0.97}_{\pm 0.01}$
0.03	$0.78_{\pm 0.09}$	$0.78_{\pm 0.10}$	$0.85_{\pm 0.05}$	$0.78_{\pm 0.09}$	$0.77_{\pm 0.10}$
0.05	$0.94_{\pm 0.02}$	$0.94_{\pm 0.02}$	$0.95_{\pm 0.02}$	$0.94_{\pm 0.02}$	$0.94_{\pm 0.02}$
0.07	$0.92_{\pm 0.07}$	$0.92_{\pm 0.06}$	$0.93_{\pm 0.05}$	$0.92_{\pm 0.06}$	$0.92_{\pm 0.07}$
0.09	$0.96_{\pm 0.01}$	$0.96_{\pm 0.01}$	$0.96_{\pm 0.01}$	$0.96_{\pm 0.01}$	$0.96_{\pm 0.01}$
Sup.	$0.96_{\pm 0.01}$	$0.96_{\pm 0.00}$	$0.96_{\pm 0.01}$	$0.96_{\pm 0.01}$	$0.96_{\pm 0.01}$
RF *0.01*	$0.87_{\pm 0.03}$	$0.87_{\pm 0.02}$	$0.88_{\pm 0.02}$	$0.87_{\pm 0.03}$	$0.87_{\pm 0.03}$
0.03	$0.92_{\pm 0.02}$	$0.93_{\pm 0.02}$	$0.92_{\pm 0.02}$	$0.92_{\pm 0.02}$	$0.92_{\pm 0.02}$
0.05	$0.90_{\pm 0.02}$	$0.90_{\pm 0.01}$	$0.90_{\pm 0.02}$	$0.90_{\pm 0.02}$	$0.90_{\pm 0.02}$
0.07	$0.90_{\pm 0.03}$	$0.91_{\pm 0.01}$	$0.91_{\pm 0.02}$	$0.90_{\pm 0.03}$	$0.90_{\pm 0.03}$
0.09	$0.95_{\pm 0.01}$	$0.96_{\pm 0.01}$	$0.95_{\pm 0.01}$	$0.95_{\pm 0.02}$	$0.95_{\pm 0.02}$
Sup.	$0.97_{\pm 0.00}$	$0.97_{\pm 0.00}$	$0.97_{\pm 0.00}$	$0.97_{\pm 0.00}$	$0.97_{\pm 0.00}$
DT *0.01*	$0.81_{\pm 0.05}$	$0.82_{\pm 0.05}$	$0.82_{\pm 0.05}$	$0.81_{\pm 0.06}$	$0.81_{\pm 0.06}$
0.03	$0.88_{\pm 0.04}$	$0.88_{\pm 0.04}$	$0.88_{\pm 0.04}$	$0.88_{\pm 0.04}$	$0.88_{\pm 0.04}$
0.05	$0.91_{\pm 0.01}$	$0.92_{\pm 0.01}$	$0.92_{\pm 0.01}$	$0.91_{\pm 0.01}$	$0.91_{\pm 0.01}$
0.07	$0.90_{\pm 0.02}$	$0.91_{\pm 0.01}$	$0.90_{\pm 0.02}$	$0.90_{\pm 0.02}$	$0.90_{\pm 0.02}$
0.09	$0.93_{\pm 0.01}$	$0.93_{\pm 0.01}$	$0.93_{\pm 0.01}$	$0.93_{\pm 0.01}$	$0.93_{\pm 0.01}$
Sup.	$0.95_{\pm 0.01}$	$0.97_{\pm 0.00}$	$0.95_{\pm 0.01}$	$0.95_{\pm 0.01}$	$0.95_{\pm 0.01}$

C. Baseline Comparison

To assess the quality of our results, we compared them with previous IG fake and bot account detection mechanisms [2], [39], [42], [48]. Only Akyon et al. [2] released their data, so we used their dataset comprising authentic and fake/bot accounts to train all the baselines, adapting the features and re-implementing the models. Each baseline was tested on all our CT accounts (provider by provider) and real accounts. Table IV reports the avg±std in detecting CT profiles for each provider. Our algorithm outperforms all the baselines, being statistically better[9] than the best baselines for Recall (p-value < 0.05) and F1-score (p-value < 0.01). The lower baselines' recall can be explained by CT accounts resembling real accounts characteristics, avoiding detection as expected. However, the relatively high standard deviations imply that the quality of CT providers varies significantly, i.e., some of them deliver lower-quality accounts, detectable by previous methods. The presence of low-quality profiles also highlighted in Table I, allowed our detector to spot both CT and *classic* fake accounts, making it more reliable than previous models trained on simple bots or synthetic data.

TABLE IV: Baseline comparison in detecting CT profiles.

Baseline	Precision	Recall	F1-Score
Thejas et al. [48]	$0.77_{\pm 0.05}$	$0.89_{\pm 0.09}$	$0.82_{\pm 0.06}$
Sheika et al. [42]	$0.94_{\pm 0.02}$	$0.84_{\pm 0.11}$	$0.88_{\pm 0.07}$
Akyon et al. [2]	$0.87_{\pm 0.05}$	$0.83_{\pm 0.19}$	$0.84_{\pm 0.14}$
Purba et al. [39]	$0.92_{\pm 0.03}$	$0.80_{\pm 0.14}$	$0.85_{\pm 0.09}$
Our	$\mathbf{0.95}_{\pm 0.02}$	$\mathbf{0.95}_{\pm 0.03}$	$\mathbf{0.95}_{\pm 0.02}$

We now explore the features' importance to explain why baselines performed worse. Figure 3 shows our model coefficients based on standardized features, so they are comparable. Baselines' most predictive features were the number of posts,

[9]Unpaired t test with $\alpha = 0.05$ as significance threshold.

following, followers, and bio length [2], [39], [42], [48]. While the number of following and posts is also crucial for us, the followers and bio length are less influential. The reason is that bots and simple fake accounts tend to have few followers and no bio, thus biasing baselines. Instead, CT profiles usually have many followers and genuine bios since they are real people profiles. Moreover, baselines do not leverage username and fullname characteristics, the number of videos, and if an account is private or verified, which are relevant to us. This suggests our model performs better due to the training data that includes CT profiles and the features we extracted (e.g., # digits in username) rather than the model itself. Nonetheless, we contribute to the state-of-the-art by demonstrating that (i) *classic* fake accounts detectors are not enough to effectively detect CT profiles, (ii) the training data are more important than the detection algorithm, (iii) the task can be efficiently solved with SSL algorithms, significantly reducing (99% less!) the time and costs to label data.
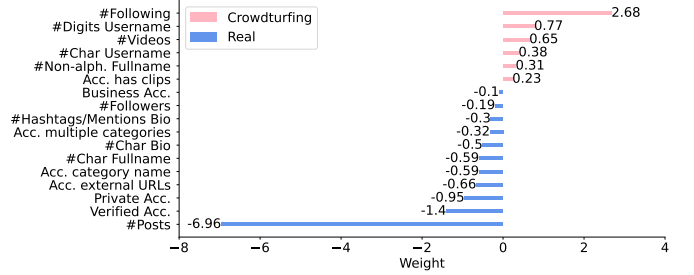

Fig. 3: Logistic Regression weights to discriminate Crowdturfing (positive label) vs Real (negative label) profiles.

V. CROWDTURFING ANALYSIS: PROFILES INFORMATION

With our CT profile detector trained, we are ready to analyze CT engagement in the wild. In our detection strategy, we detect profiles involved in CT using our model (§IV-B) and label their engagement accordingly. Since CT profiles contribute to a fake engagement, we will also refer to them as *fake* (non-genuine) accounts and engagement vs. *real* (genuine) ones. For our analyses, we collected the comments and commenters' profile information[10] of 50 recent posts of 20 mega-influencers with over 1 million followers (1000 posts in total). We selected posts at least five days old to allow IG automatically remove *classic* fake interactions [28], [29]. The influencers come from different nationalities and the following categories: fashion, beauty, fitness, art, music, lifestyle, and family. In total, we gathered 603,007 comments generated by 248,388 unique users. The reasons why we collected only comments-related information and e.g., not likes, are discussed in the comments analysis section (§VI).

Our CT detection model detected 55,719 CT profiles among the 248,388 collected ($\sim$22%). This percentage aligns with the estimate of 20-40% in celebrities' accounts [10]. We acknowledge that some of the detected accounts may not be CT; however, we are still dealing with "advanced" fake profiles

[10]Profiles info were collected via Instaloader https://instaloader.github.io/.

that have bypassed (i) the automatic screening mechanisms employed by IG [28], [29] (and Meta [6] in general), and (ii) the potential moderation done by the influencers themselves (e.g., by removing blatant spam comments). Therefore, we can assume our further analyses will primarily focus on CT or advanced fake profiles that resemble and act as legitimate profiles. In this section, we provide a detailed study of CT profiles' information, including the number of followers and following (§V-A), biography (§V-B), and external URLs (§V-C), to determine whether CT profiles engage in malicious activities besides crowdturfing.

A. Followers and following ratio analysis

To increase other accounts' engagement (and therefore gain more money), a fake account will display an unusually high number of following (§III). Conversely, genuine users should have a more balanced ratio of followers and following according to IG averages [26]. Figure 4 shows the mean and std of followers/following for fake and real users. Real users are divided into normal and influencers tiers[11] as follows:

- *Normal accounts*: less than 1,000 followers;
- *Nano influencers*: $1,000 \leq$ followers $< 10,000$;
- *Micro influencers*: $10,000 \leq$ followers $< 50,000$;
- *Mid-tier influencers*: $50,000 \leq$ followers $< 500,000$;
- *Macro influencers*: $500,000 \leq$ followers $< 1,000,000$;
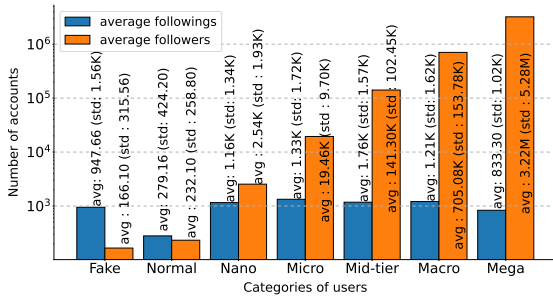- *Mega influencers*: more than 1,000,000 followers.

Fig. 4: Followers and following avg and std of CT users (Fake) and different categories of real users. Y-axis is in log scale.

The graph shows that the number of followers of fake users is (on average) much smaller than the number of following. Indeed, these accounts are incentivized to follow more people to grow their earnings through CT activities, confirming our initial assumption. Following and followers of normal users are balanced, but as the popularity of the genuine account grows, followers increase exponentially while following hovers around 1000. For more popular influencers, the standard deviation increases simply because their categories include wider ranges (e.g., from one to hundreds of million followers for mega influencers). We further inspected the following distribution of CT accounts in Figure 5. Most CT accounts have between 0 and 500 following, with the number decreasing as the following increases, suggesting CT accounts tend to
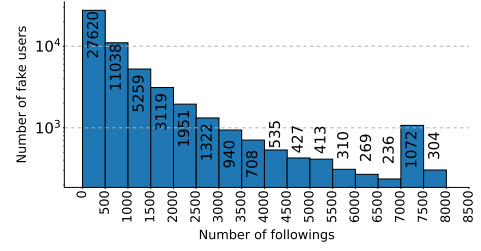
Fig. 5: Distribution of fake accounts' following.

maintain a low profile to avoid being flagged as spammers. An exception occurs in the last two bins. IG introduced a 7500 following limit[12] to contrast spamming activities, and many CT (probably more similar to *classic* fake) accounts are just below this limit. Despite it, 304 fake profiles likely surpassed the threshold before its introduction.

B. Fake profiles biography analysis

Many *classic* fake IG accounts use a catchy biography to lure victims into clicking malicious links. Thus, we tried to find suspicious words in the CT users' biographies. To this aim, we created a list of 31 elements, including words and emojis often used by this fake user. The list, based on our knowledge of fake behavior and a brief manual inspection, contained words like "stories", "chat", "follow", "gain", "click", "link", and emojis usually linked to malicious or sexual activities, like "🔞", "⬇️", "🤤", "💦", "💋" [11]. Only 5635 CT accounts (10.11% of the total detected) had at least one of the elements of the list. Thus, most CT accounts do not seek to boost their profiles or induce people to click links. Rather, they are interested in making profits by increasing the engagement of other accounts.

C. Fake profiles external URLs analysis

The last analysis performed on the CT accounts is based on their external URLs, aiming to understand the most used URLs among CT users and whether they could be vectors of attacks conducted over social networks [35]. Of the total fake accounts, only 2834 (5.08%) had an external URL on their profile page. We grouped these 2834 URLs into the following categories:

- *Videogame*: Youtube, Twitch, Discord;
- *Messaging*: WhatsApp, Telegram;
- *Social Network*: Facebook, Twitter, Instagram, etc.;
- *Music & Photography*: Spotify, Soundcloud, Vsco.co;
- *Email & Google services*: Gmail, Maps, Outlook;
- *URL redirecting*: Linktr.ee, Tinyurl, Linkr.bio, Bit.ly;
- *Shopping & Payment*: PayPal, Vinted, Amazon, etc.;
- *Personal website & Petition*: Blogspot, Wordpress, etc.;
- *Adult content*: URLs to different adult websites;
- *Other*.

Inside the categories, we also included shortened URLs (e.g., wa.me or t.me for WhatsApp and Telegram, respectively). The results are shown in Figure 6. Remark that

---

[11]https://www.shopify.com/id/blog/instagram-influencer-marketing

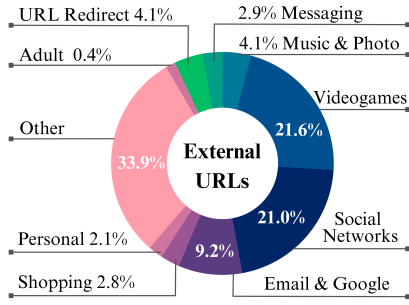[12]https://help.instagram.com/408167069251249

Fig. 6: Categories of External URLs of the fake profiles.

even if the categories contain well-known websites, some can be used for malicious purposes. For instance, we found many WhatsApp links starting a conversation or a phone call with strangers who could easily be scammers. Similarly, we inspected and monitored Telegram URLs, grouping them into:

- *Conversation*: Similarly to WhatsApp URLs, starts a conversation with a potential scammer;
- *Piracy*: Illegal groups that share movies and tv series;
- *Selling*: Scam groups that try to sell clothes, Amazon gift cards, cryptocurrencies, NFTs, etc.

Moreover, *classic* fake profiles commonly use redirect URLs to route the victim to a malicious site [31]. From Figure 6, it is possible to see that the "Other" section is more relevant than the other categories inside the pie chart, with precisely 961 URLs. It contains very heterogeneous URLs, making their categorization challenging. To better understand these URLs' nature (i.e., if they are malicious), we have relied on a fraud prevention and detection service called Ipqualityscore.[13] It allows checking for suspicious links by using a mixture of blacklists and deep learning algorithms, and to define the following URLs categories:

- *Parked*: Domains that have been dormant for a long time;
- *Spamming*: Websites that spams malicious content;
- *Malware*: Websites hosting viruses, malware, etc.;
- *Phishing*: Websites hosting fake login, or sign up forms;
- *Adult*: Websites that contain adult content.

The results of this evaluation are shown in Figure 7. For convenience, we grouped the "Phishing", "Malware", and "Adult" categories since they had very few matches. From the total 961 "Other" URLs, 599 were considered safe, while the remaining 362 were divided as follows:

- 190 URLs were parked and/or spamming websites;
- 5 URLs were marked as malware websites;
- 7 URLs were marked as phishing websites;
- 11 URLs were adult websites;
- 149 were considered suspicious websites.

These results show that most external URLs in the "Other" category were considered safe. However, many spamming and suspicious websites can be used for malicious purposes. Comparing the obtained results to the overall number of CT users, we can confirm that most are solely involved in CT activities rather than malicious activities.
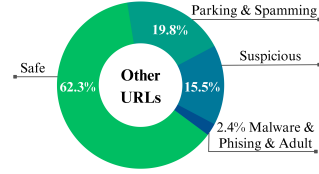
[13]https://www.ipqualityscore.com/

Fig. 7: Results provided by the fraud prevention and detection service on the URLs in the "Other" category.

VI. REAL VS CROWDTURFING COMMENTS ANALYSIS

This section analyses CT engagement. In particular, we aim to understand if CT can be directly spotted by actions (e.g., comments) instead of leveraging profile information. As stated before, CT profiles are driven by humans, so intuitively, there should be little to no difference between real and fake engagement, but we cannot draw conclusions without proper analysis. On IG, the primary forms of engagement are liking and commenting. CT likes cannot be isolated from the action itself since it carries no information beyond temporal data (unavailable on IG). Instead, comments provide valuable information (e.g., stylometric features) that could be used for CT detection. Moreover, comments present a higher level of public expression than likes [4] and are considered more important to boost the visibility of an account [8], [46]. For these reasons, we focus on comments in this section, presenting five studies to spot the differences between comments made by CT and real users.

A. Stylometric Analysis

From our dataset, we isolated 121,822 comments shared by CT users and 481,185 from legit ones. We performed a stylometric analysis similar to the one conducted in [7], based on Lexical Features, Syntactical Features, and Emoji Features.

Lexical Features. We calculated the number of sentences per comment, the number of words in each comment, the number of words in each sentence, and the length of the comments. We found several statistically significant (p-value < 0.001) differences: CT users have an overall mean of 1.13 words per comment, while the real ones have 4.34. Similarly, the number of words per sentence is 0.94 for the CT accounts and 2.96 for the real ones. Instead, both categories of users have a mean of 1.35 sentences per comment. In each comment, there is a low repetition of words: we obtained that 99% of them, made by CT users, have no word repetitions, while for the legit users is 97%. Another important distinction is the length of the comments: the CT users shared text with a mean length of 28.89 (std: 61.19) characters (emojis included), while the legit users have a mean of 23.74 (std: 46.74). Even if similar, they are statistically significant (p-value < 0.001). The emoji comparison better explains how real users, with more words, have shorter comments.

Syntactical Features. We counted the number of comments starting with a capital letter, punctuation present in the text, and capital words. We found very close results between CT and real users: the beginning of the comment is in uppercase

for 33.86% of comments made by CT users and for 34.94% of real ones. 35% of the comments have some punctuation for both accounts categories. Finally, we saw that both categories do not use upper-cased words: the mean of the ratios between uppercase words and all the words in each comment are 0.021 for the CT users and 0.025 for the legit ones.

Emoji Features. We detected emojis in the comments using demoji[14]. Our study focused on the presence of emojis and alphanumerical text in the comments, in particular:

1) **The percentage of comments with at least one emoji**;
2) **Most used emojis**: the percentage of an emoji among all the fake comments. Multiple occurrences of the same emoji on the same comment increase the counter by one.
3) **Avg emojis when present**: considering only comments presenting emojis, the avg number of them. Multiple occurrences of the same emoji increase the counter accordingly.

Table V reports the results. The top-most emojis used are equal for both users, with similar percentages. Another meaningful result is that even if real users have, on average, slightly more comments with emojis, the quantity of emojis in such comments is fewer compared to CT users. This result might explain the outcomes on comment length found in §VI-A. To sum up, results obtained so far show some stylometric differences, but mostly similarities between CT and real users when the focus is on emoji used, sentences per comments, or syntactical features. Legit users share comments with more words, fewer emojis, and an overall shorter comment length.

TABLE V: Emojy-based Stylometric analysis. CE = Comments with Emoji, EPC = Avg Emoji per comment.

	CE (%)	% of Most used Emoji						EPC
		❤️	😍	🔥	👏	😂	🙌	
Fake	71.6	25.18	19.92	10.57	4.91	4.03	2.73	3.557
Real	72.7	22.30	18.46	14.42	5.00	4.92	3.04	3.211

B. Common Words Analysis

We analyzed the most common words CT and real profiles use. As a pre-processing, we removed emojis, punctuations, and unproductive words with less than three characters, e.g., "and", "the", "you". The word clouds in Figure 8 show fake and real users' top 100 most used words. In general, we found a lot of positive and loving expressions, such as beautiful, love, happiness, niceness, etc.

(a) Fake Users (b) Real Users

Fig. 8: Most used words by fake and real users.

An interesting word from Figure 8a is "Dokter", which appeared in 1069 comments. By investigating the accounts

[14]https://pypi.org/project/demoji/

spamming this word, we might have found a botnet whose objective is to spam "IG doctors" accounts. All these doctors' profiles have a WhatsApp business link starting a chat with a message to complete: "*NAME*: *CITY/STATE*: *ORDER/COMPLAINTS*: *AGE*:". Some doctors' accounts no longer exist on IG, suggesting they probably violated the ToS. Other similar accounts had the format "dr.[doctor_name]", presenting the same WhatsApp link and conversation, but different phone numbers. We found 1370 comments coming from 33 different accounts containing such words, suggesting the presence of a bigger malicious network.

C. Number of Comments per User

248,388 unique users posted the 603,007 comments we analyzed; thus, many users posted multiple comments. We found that a legit user, on average, has posted 1.95 comments (std 5.94), while a CT user has posted slightly more (2.24, std 7.57). The result obtained in this analysis complies with the one in §VI-A: a CT user has a similar behavior as the legit user. However, a CT user generally shares more comments than a real one because their purpose is to generate engagement. But to avoid IG bot detection, an account has to act like a real human being.

D. Language Analysis

We analyzed the language used by CT and real users using SpaCy.[15] The text was filtered out of emojis and then used as input for the neural network. The results are shown in Figure 9. In both CT and real comments, we found that the prominent language is English (35.2% and 43.5%, respectively), followed by Japanese and French. The "Other" slices include more than 100 languages, each with a presence below 2%. They are probably the second largest sections of the pie charts because many comments just mentioned other accounts or used single words, complicating the language detection process. Besides that, CT users likely adopt the language of their target community or, more commonly, English. In fact, as stated in §III, many CT providers allow the option to deliver followers from specific geographical locations.
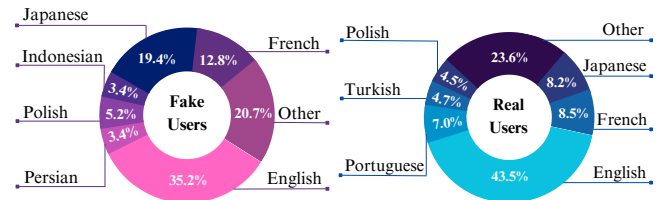

Fig. 9: Languages detected in comments.

E. Topics Analysis

To further investigate the behavior of CT and real users, we inspected the topics in their comments. Many state-of-the-art topic modeling algorithms, such as Latent Dirichlet Allocation (LDA), require long text to extract topics. However, social network comments are usually concise sentences, making

[15]https://spacy.io/usage/facts-figures

TABLE VI: Top-10 topics extracted from fake and real comments.

Fake Comments			Real Comments		
N. Comm.	*Top Words*	*Label*	*N. Comm*	*Top Words*	*Label*
3817	beautiful gorgeous sexy perfect hot amaze girl	Female Beauty	13034	beautiful gorgeous nice cute pretty lovely girl	Female Beauty
2290	love beautiful cute smile god woman world girl	Love (woman)	7547	love good smile congrats great brother bro wish	Love (Males)
2117	good want video well thank man bro life bike work	Man Compliment	6983	dream make want come time good life day hope	Life Dreams
1755	happy birthday halloween republic thanksgiving	Pagan holidays	6274	christmas merry god bless family thank bible	Christmas
1476	please christmas merry story follow check thank	Christmas/Follow	6035	happy new year birthday day family love republic	Pagan holidays
1136	help fire turkey people stop please give helpturkey	Help Turkey	6008	help need fire people turkey please animal world	Turkey/Ecologists
1021	trop wanna kiss lip red face belle pretty liplock	Kiss & Face	5658	picture crazy bro think top video sick bike man	Exalt Men
674	arm chest belly waist neck armpit thigh dance	Body parts	5524	follow check story post page like support profile	Follow/support
514	problem solution wife money call whatsapp expert	Problems/Ads	4846	please congrats reply check story dance song real	Music
223	love back help massage oil bubbs real magic	Relax	1381	problem belle family life help solution marriage	Family Problems

the topic modeling more challenging. In our experiments, we used GPU-PDMM [33], which is typically adopted to extract topics of tweets. Based on the Poisson-based Dirichlet Multinomial Mixture (PDMM) model, GPU-PDMM promotes the semantically related words under the same topic during the sampling process by using the Generalized Polya Urn (GPU) model. We considered only English comments for the analysis, after removing non-alphabetical characters, emojis, stop words, words shorter than three characters, and applying lemmatization. From our comments, 15,023 CT comments and 63,290 Real comments were suitable for the study. We instructed the model to distinguish ten topics in an unsupervised fashion, returning for each comment the belonging topic and the top words associated with each topic. The results of the topics inference are shown in Table VI.

As expected, we find high alignment between topics covered by CT and real profiles. Most comments exalt female beauty, using compliments, love words, or positive feelings to boost engagement. In particular, CT comments contain more exaggerated terms, such as "sexy", "perfect", or "amaze". Conversely, we found few advertisement comments, likely to avoid being flagged as spammers. An interesting difference between CT and real comments is how they dealt with the *Help Turkey* topic. For real profiles, we found additional words such as "animal" and "world", suggesting they also brought up other environmental arguments, while CT did not. For real comments, we also found a *Follow/support* topic, which could be a false positive (some spammers were not detected) or that they did not care about being labeled as spammers. In summary, the topic analysis revealed some differences, but not consistently enough to allow for proper differentiation.

VII. CONCLUSION

In this work, we developed an algorithm that leverages profiles' characteristics through semi-supervised learning to spot IG crowdturfing activities. To train our classifier, we purchased CT profiles from 11 providers, which we further studied to understand their services and the type of profiles involved in them. Our Logistic Regression classifier scored 0.95% F1-score. To spot IG CT activities in the wild, we targeted the most recent posts of 20 influencers of different nationalities and categories. We mainly focused on comments, as they are a crucial engagement metric for accounts' visibility, and carry more information than likes. For this purpose, we collected 603,007 comments among the different posts made by 248,388 unique users. Our model labeled 55,719 of these profiles as CT accounts. We compared CT profiles and comments with genuine ones, concluding that CT activities would be difficult to detect based only on their activities. Indeed, CT profiles are mostly real profiles guided by real humans; thus, their activities are close to genuine ones. In contrast to bots or fake profiles, they seem to not be involved with malicious activities besides boosting other accounts' engagement. In the future, we plan to distinguish between CT profiles and other "advanced" fake profiles we might have (in)voluntarily encountered in our analyses. While IG and the research community focused a lot on detecting bots and automated accounts, we believe more studies should be conducted on CT activities or in general, advanced fake profiles which negatively impact influencer marketing, IG, and most of its users.

ETHICAL CONSIDERATIONS

We faced two main ethical challenges: CT activities' involvement and data collection on IG. Our experiments were designed following the exemption guideline from a formal review by our institute's IRB. To deal with CT activities, we acted similarly to previous works that analyze underground activities [44], [49], [50], first by dealing only with a small number of CT followers and platforms, minimizing our effect on them and IG. Second, we linked the followers to freshly created accounts that had no prior connection with other IG accounts, and we deleted them at the end of the study. Thus, CT activities were not involving legitimate users.

For data collection, we gathered only profiles' information and comments publicly available, removing all the information linked to individual subjects (e.g., name, profile picture). Similar to previous works [25], [40], we could not request informed consent to prevent participants from (in)voluntarily changing their behavior, causing the Hawthorne effect [22]. Since IG APIs do not return all the public information of a user's profile, yet visible by simply browsing it, we collected such data in an automated way, which is not allowed by the ToS. However, as argued by Fiesler et al. [21], "ethical decisions regarding data collection should go beyond ToS and consider contextual factors of the source and research". In particular, IG ToS allows manual collection, suggesting that automated collection is probably not allowed to avoid heavy servers' workload [21]. Therefore, we tuned our tools to collect data at a slow human-like pace, using only our 11 profiles over five months, avoiding any ban from the platform.

REFERENCES

[1] T. Abbas and U. Gadiraju. Goal-setting behavior of workers on crowdsourcing platforms: An exploratory study on mturk and prolific. In *AAAI Conference on Human Computation and Crowdsourcing*, 2022.

[2] F. C. Akyon and M. E. Kalfaoglu. Instagram fake and automated account detection. In *IEEE ASYU*, 2019.

[3] M. Al-Qurishi et al. Sybiltrap: A graph-based semi-supervised sybil defense scheme for online social networks. *Concurrency and Computation: Practice and Experience*, 30(5):e4276, 2018.

[4] K. K. Aldous et al. View, like, comment, post: Analyzing user engagement by topic at 4 levels across 5 social media platforms for 53 news organizations. In *ICWSM*, 2019.

[5] H. Alvari, E. Shaabani, and P. Shakarian. Semi-supervised causal inference for identifying pathogenic social media accounts. In *Identification of Pathogenic Social Media Accounts*, pages 51–61. Springer, 2021.

[6] G. Apruzzese, H. S. Anderson, S. Dambra, D. Freeman, F. Pierazzi, and K. A. Roundy. "Real Attackers Don't Compute Gradients": Bridging the Gap between Adversarial ML Research and Practice. In *Proceedings of the 1st IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2023.

[7] M. Bhargava, P. Mehndiratta, and K. Asawa. Stylometric analysis for authorship attribution on twitter. In *International Conference on Big Data Analytics*, pages 37–47. Springer, 2013.

[8] B. Chacon. 5 things to know about the instagram algorithm. https://later.com/blog/instagram-algorithm/, 2017. Acc. Oct 2022.

[9] A. Chetan, B. Joshi, H. S. Dutta, and T. Chakraborty. Corerank: Ranking to detect users involved in blackmarket-based collusive retweeting activities. In *Proceedings WSDM*, pages 330–338, 2019.

[10] M. Conti, J. Gathani, and P. P. Tricomi. Virtual influencers in online social media. *IEEE Communications Magazine*, 2022.

[11] M. Conti, L. Pajola, and P. P. Tricomi. Captcha attack: Turning captchas against humanity. *arXiv preprint arXiv:2201.04014*, 2022.

[12] S. Cresci. A decade of social bot detection. *Communications of the ACM*, 2020.

[13] A. Dorri, M. Abadi, and M. Dadfarnia. Socialbothunter: Botnet detection in twitter-like social networking services using semi-supervised collective classification. In *DASC/PiCom/DataCom/CyberSciTech*, 2018.

[14] H. S. Dutta, K. Aggarwal, and T. Chakraborty. Decife: Detecting collusive users in blackmarket following services on twitter. In *32nd ACM conference on hypertext and social media*, 2021.

[15] H. S. Dutta, U. Arora, and T. Chakraborty. Abome: A multi-platform data repository of artificially boosted online media entities. In *ICWSM*, 2021.

[16] H. S. Dutta and T. Chakraborty. Blackmarket-driven collusion among retweeters–analysis, detection, and characterization. *IEEE TIFS*, 2019.

[17] H. S. Dutta and T. Chakraborty. Blackmarket-driven collusion on online media: A survey. *ACM/IMS TDS*, 2(4):1–37, 2022.

[18] H. S. Dutta, N. Diwan, and T. Chakraborty. Weakening the inner strength: Spotting core collusive users in youtube blackmarket network. In *ICWSM*, volume 16, pages 147–158, 2022.

[19] H. S. Dutta, M. Jobanputra, H. Negi, and T. Chakraborty. Detecting and analyzing collusive entities on youtube. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–28, 2021.

[20] T. Elmas, R. Overdorf, A. F. Özkalay, and K. Aberer. Ephemeral astroturfing attacks: The case of fake twitter trends. In *2021 IEEE EuroS&P*, pages 403–422. IEEE, 2021.

[21] C. Fiesler, N. Beard, and B. C. Keegan. No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. In *ICWSM*, 2020.

[22] R. H. Franke and J. D. Kaul. The hawthorne experiments: First statistical interpretation. *American sociological review*, pages 623–643, 1978.

[23] W. Geyser. The state of influencer marketing 2022: Benchmark report. https://influencermarketinghub.com/influencer-marketing-benchmark-report/, 2022. Acc. Nov 2022.

[24] N. Z. Gong et al. Sybilbelief: A semi-supervised learning approach for structure-based sybil detection. *IEEE TIFS*, 2014.

[25] C. L. Hanson et al. Tweaking and tweeting: exploring twitter for nonmedical use of a psychostimulant drug (adderall) among college students. *Journal of medical Internet research*, 15(4):e2503, 2013.

[26] Hashatgsforlikes. Instagram followers: How many does the average person have? https://hashtagsforlikes.co/blog/instagram-followers-how-many-does-the-average-person-have/, 2020. Acc. Mar 2022.

[27] D. C. Hernandez-Bocanegra, A. Borchert, F. Brünker, G. K. Shahi, and B. Ross. Towards a better understanding of online influence: Differences in twitter communication between companies and influencers. *ACIS 2020 Proceedings*, 2020.

[28] Instagram. Reducing inauthentic activity on instagram. https://about.instagram.com/blog/announcements/reducing-inauthentic-activity-on-instagram, 2018. Acc. Feb 2023.

[29] Instagram. Introducing new authenticity measures on instagram. https://about.instagram.com/blog/announcements/introducing-new-authenticity-measures-on-instagram/, 2020. Acc. Feb 2023.

[30] S. Kim and J. Han. Detecting engagement bots on social influencer marketing. In *International Conference on Social Informatics*, 2020.

[31] K. Lakshmi. Bot comments on instagram are becoming horrendous. https://eatmy.news/2020/11/bot-comments-on-instagram-are-becoming.html, 2020. Acc. Mar 2022.

[32] K. Lee, S. Webb, and H. Ge. The dark side of micro-task marketplaces: Characterizing fiverr and automatically detecting crowdturfing. In *ICWSM*, 2014.

[33] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma. Enhancing topic modeling for short texts with auxiliary word embeddings. 2016.

[34] X. Liao et al. Should we trust influencers on social networks? on instagram sponsored post analysis. In *ICCCN*, 2021.

[35] W. Luo, J. Liu, J. Liu, and C. Fan. An analysis of security in social networks. In *2009 IEEE DASC*, 2009.

[36] L. Manikonda, V. V. Meduri, and S. Kambhampati. Tweeting the mind and instagramming the heart: Exploring differentiated content sharing on social media. In *ICWSM*, 2016.

[37] E. Morales. Instagram bots in 2022 — the best bots and everything else you need to know. https://bettermarketing.pub/instagram-bots-in-2021-everything-you-need-to-know-b57fb0a3b8e9, 2021. Acc. 03-28-2022.

[38] M. Orabi et al. Detection of bots in social media: A systematic review. *Information Processing & Management*, 57(4):102250, 2020.

[39] K. R. Purba, D. Asirvatham, and R. K. Murugesan. Classification of instagram fake users using supervised machine learning algorithms. *IJECE*, 10(3):2763, 2020.

[40] D. Quercia, J. Ellis, L. Capra, and J. Crowcroft. In the mood for being influential on twitter. In *2011 IEEE PST/SCSM*, 2011.

[41] P. K. Roy and S. Chahar. Fake profile detection on social networking websites: a comprehensive review. *IEEE TAI*, 2020.

[42] S. Sheikhi. An efficient method for detection of fake accounts on the instagram platform. *Rev. d'Intelligence Artif.*, 34(4):429–436, 2020.

[43] N. Shi et al. Semi-supervised random forest for intrusion detection network. In *MAICS*, pages 181–185, 2017.

[44] J. Song, S. Lee, and J. Kim. Crowdtarget: Target-based detection of crowdturfing in online social networks. In *CCS*, 2015.

[45] Statista. Most popular social networks worldwide as of january 2022. https://statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. Acc. Oct 2022.

[46] Statusbrew. Instagram algorithm 2022: How to conquer it. https://statusbrew.com/insights/instagram-algorithm/, 2022. Acc. Sep 2021.

[47] S. Stieglitz, F. Brachten, B. Ross, and A. Jung. Do social bots dream of electric sheep? a categorisation of social media bot accounts. In *27th Australian Conference on Information Systems*, pages 1–11, 2018.

[48] G. Thejas et al. Learning-based model to fight against fake like clicks on instagram posts. In *2019 SoutheastCon*, pages 1–8. IEEE, 2019.

[49] K. Thomas, D. McCoy, C. Grier, A. Kolcz, and V. Paxson. {Trafficking} fraudulent accounts: The role of the underground market in twitter spam and abuse. In *USENIX Security*, pages 195–210, 2013.

[50] G. Voronin, A. Baumann, and S. Lessmann. Crowdturfing on instagram-the influence of profile characteristics on the engagement of others. In *Twenty-Sixth European Conference on Information Systems 2016*, 2018.

[51] G. Wang, C. Wilson, X. Zhao, Y. Zhu, M. Mohanlal, H. Zheng, and B. Y. Zhao. Serf and turf: crowdturfing for fun and profit. In *21st WWW*, pages 679–688, 2012.

[52] J. Wise. How much time do people spend on social media 2022? https://earthweb.com/how-much-time-do-people-spend-on-social-media/, 2022. Acc. Nov 2022.

[53] X. Yang, Q. Yang, and C. Wilson. Penny for your thoughts: Searching for the 50 cent party on sina weibo. In *ICWSM*, 2015.

[54] K. Zarei et al. How impersonators exploit instagram to generate fake engagement? In *IEEE ICC*, 2020.