Social Honeypot for Humans: Luring People Through Self-managed Instagram Pages

Sara Bardi[1], Mauro Conti[1,2], Luca Pajola[1], and Pier Paolo Tricomi[1,2](✉)

[1] University of Padua, Padua, Italy
sara.bardi@studenti.unipd.it, {conti,pajola,tricomi}@math.unipd.it
[2] Chisito S.r.l, Padua, Italy

Abstract. Social Honeypots are tools deployed in Online Social Networks (OSN) to attract malevolent activities performed by spammers and bots. To this end, their content is designed to be of maximum interest to malicious users. However, by choosing an appropriate content topic, this attractive mechanism could be extended to *any* OSN users, rather than only luring malicious actors. As a result, honeypots can be used to attract individuals interested in a wide range of topics, from sports and hobbies to more sensitive subjects like political views and conspiracies. With all these individuals gathered in one place, honeypot owners can conduct many analyses, from social to marketing studies.

In this work, we introduce a novel concept of social honeypot for attracting OSN users interested in a generic target topic. We propose a framework based on fully-automated content generation strategies and engagement plans to mimic legit Instagram pages. To validate our framework, we created 21 self-managed social honeypots (i.e., pages) on Instagram, covering three topics, four content generation strategies, and three engaging plans. In nine weeks, our honeypots gathered a total of 753 followers, 5387 comments, and 15739 likes. These results demonstrate the validity of our approach, and through statistical analysis, we examine the characteristics of effective social honeypots.

Keywords: Social Networks · Social Honeypots · Instagram · User Profiling · Artificial Intelligence · Privacy

1 Introduction

In recent years, Social Network Analysis (SNA) has emerged as a powerful tool for studying society. The large amount of relational data produced by Online Social Networks (OSN) has greatly accelerated studies in many fields, including modern sociology [62], biology [23], communication studies [25], and political science [36]. SNA success can be attributed to the exponential growth and popularity OSN faced [4], with major OSN like Facebook and Instagram (IG) having billions of users [35,58]. Researchers developed a variety of tools for SNA [56]; however, elaborating the quintillion bytes of data generated every day [30] is

M. Tibouchi and X. Wang (Eds.): ACNS 2023, LNCS 13905, pp. 309–336, 2023.
https://doi.org/10.1007/978-3-031-33488-7_12

far from trivial [9]. The computational limitations compel scientists to conduct studies on sub-samples of the population, often introducing bias and reducing the quality of the results [8]. Furthermore, the reliability of data is hindered by adversarial activities perpetuated over OSN [12,33], such as the creation of fake profiles [60], crowdturfing campaigns [69,71], or spamming [28,50,80].

Back in the years, cybersecurity researchers proposed an innovative approach to overcome the computational limitation in finding malicious activity in OSN (e.g., spamming), by proposing social honeypots [41,66,73]: profiles or pages created ad-hoc to lure adversarial users, analyze their characteristics and behavior, and develop appropriate countermeasures. Thus, their search paradigm in OSN shifted from "look for a needle in the haystack" (i.e., searching for spammers among billions of legit users) to "the finer the bait, the shorter the wait" (i.e., let spammers come to you).

Motivation. The high results achieved by such techniques inspired us to generalize the approach, gathering in a *single* place *any target users* we wish to study. Such a framework's uses are various, from the academic to the industrial world. First, *profilation* or *marketing* toward target topics: IG itself provides page owners to know aggregated statistics (e.g., demographic) of their followers and users that generate engagement.[1] Second, *social cybersecurity analytics*: researchers or police might deploy social honeypots on sensitive themes to attract and analyze the behavior of people who engage with them. Examples of themes are fake news and extremism (e.g., terrorism). Although our "general" social honeypot may be used either benignly (e.g., to find misinformers) or maliciously (e.g., to find vulnerable people to scam), in this paper, we only aim to examine the feasibility of such a tool, and its effectiveness. Moreover, we investigate whether this technique can be fully automated, limiting the significant effort of creating a popular IG page [59]. We focus on IG given its broad audience and popularity. Furthermore, IG is the most used social network for marketing purposes, with nearly 70 percent of brands using IG influencers (even virtual [11]) for their marketing campaigns [29].

Contribution. In this work, we present an automated framework to attract and collect legitimate people in social honeypots. To this aim, we developed several strategies to understand and propose guidelines for building effective social honeypots. Such strategies consider both *how to generate content automatically* (from simple to advanced techniques), and *how to engage with the OSN* (from naive to complex interactions). In detail, we deployed 21 honeypots and maintained them for nine weeks. Our four content generation strategies involve state-of-the-art Deep Learning techniques, and we actively engage with the network following three engagement plans.

[1] Instagram API provides to the owner aggregated statistics of followers (gender, age, countries) when their page reaches 100 followers [18].

The main contributions of our paper can be summarized as follows:

- We define a novel concept of Social Honeypot, i.e., a flexible tool to gather *real people* on IG interested in a target topic, in contrast to previous studies focusing on malicious users or bots;
- We propose four automatic content generation strategies and three engagement plans to build self-maintained IG pages;
- We demonstrate the quality of our proposal by analyzing our 21 IG social honeypots after a nine weeks period.

Outline. We begin our work discussing related works (§2). Then, we present our methodology and implementation in §3 and §4. In §5, we evaluate the effectiveness of our honeypots, while §6 presents social analyses. We discuss the use cases of our approach and its challenges in §7 and conclude the paper in §8.

2 Related Works

Honeypot. Honeypots are decoy systems that are designed to lure potential attackers away from critical systems [64]. Keeping attackers in the honeypot long enough allows to collect information about their activities and respond appropriately to the attack. Since legit users have no valid reason to interact with honeypots, any attempt to communicate with them will probably be an attack. Server-side honeypots are mainly implemented to understand network and web attacks [34], to collect malware and malicious requests [76], or to build network intrusion detection systems [37]. Conversely, client-side honeypots serve primarily as a detection tool for compromised (web) servers [49,72].

Social Honeypot. Today, honeypots are not limited to fare against network attacks. Social honeypots aim to lure users or bots involved in illegal or malicious activities perpetuated on Online Social Networks (OSN). Most of the literature focused on detecting spamming activity, i.e., unsolicited messages sent for purposes such as advertising, phishing, or sharing undesired content [66]. The first social honeypot was deployed by Webb et al. [73] on MySpace. They developed multiple identical honeypots operated in several geographical areas to characterize spammers' behavior, defining five categories of spammers. Such work was extended to Twitter by Lee et al. in 2010 [41], identifying five more spammers' categories, and proposing an automatic tool to distinguish between spammers and legit users. Stringhini et al. [66] proposed a similar work on Facebook, using fake profiles as social honeypots. Similarly to previous works, these profiles were passive, i.e., they just accepted incoming friend requests. Their analysis showed that most spam bots follow identifiable patterns, and only a few of them act stealthily. De Cristofaro et al. [15] investigated Facebook Like Farms using social honeypots, i.e., blank Facebook pages. In their work, they leveraged demographic, temporal, and social characteristics of likers to distinguish between genuine and fake engagement. The first "active" social honeypot was developed on Twitter by Lee et al. [42], tempting, profiling, and filtering content polluters

in social media. These social honeypots were designed to not interfere with legitimate users' activities, and learned patterns to discriminate polluters and legit profiles effectively. 60 honeypots online for seven months gathered 36'000 interactions. More active social honeypots were designed by Yang et al. [75]), to provide guidelines for building effective social honeypots for spammers. 96 honeypots online for five months attracted 1512 accounts. Last, pseudo-honeypots were proposed by Zhang et al. [79], which leveraged already popular Twitter users to attract spammers efficiently. They run 1000 honeypots for three weeks, reaching approximately 54'000 spammers.

Differences with Previous Work. To date, social honeypots have been mainly adopted to detect spammers or bot activities. The majority of research focused on Twitter, and only a few works used other social networks like Facebook. There are several reasons behind this trend. First, spamming is one of the most widespread malicious activities on social networks because it can lead to other more dangerous activities. Second, Twitter APIs and policies facilitate data collection, and there are widely adopted Twitter datasets that can be used for further analysis. To the best of our knowledge, there are no works that utilize social honeypots on Instagram, perhaps because it is difficult to distribute, maintain and record honeypots' activities on this social network. Moreover, our goal is to attract *legit users* rather than spammers, which is radically different from what was done insofar. Indeed, many analyses could be easier to conduct by gathering people in one place (e.g., an IG page). For instance, a honeypot could deal with peculiar topics to simplify community detection [7], could advertise a product to grasp consumer reactions [10], understand political views [45], analyze and contrast misinformation [16], conspiracies [2], and in general, carry out any Social Network Analytics task [19]. Last, owners of IG pages can see the demographic information of their followers (inaccessible otherwise), having extremely helpful (or dangerous) information for further social or marketing analyses [61].

3 Methodology

3.1 Overview and Motivation

The purpose of our social honeypots is to attract people interested in a target topic. The methodology described in this section is intended for Instagram (IG) pages, but it can be extended to any generic social network (e.g., Facebook) with minor adjustments. We define the social honeypot as a combination of three distinct components: (i) the honeypot *topic* that defines the theme of the IG page (§3.2); (ii) the *generation strategy* for creating posts related to a target topic (§3.3); (iii) the *engagement plan* that describes how the honeypot will engage the rest of the social network (§3.4). Figure 1 depicts the social honeypot pipeline.

Our study examines different types of honeypots with a variety of topics, generation strategies, and engagement plans, outlined in the rest of this section. Our experiments aim to answer the following research questions:

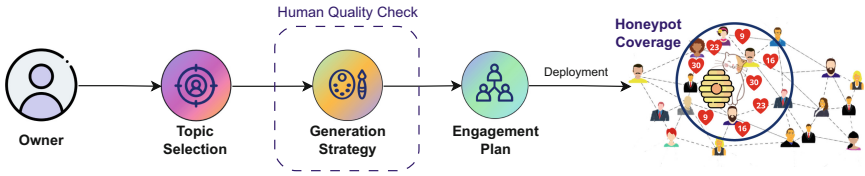

Fig. 1. Pipeline overview to create a social honeypot. After the owner decides on the topic, generation strategy, and engagement plan, the honeypot automatically generates posts to interact with the social network. After the post is automatically generated, the owner can approve it or request a new one to meet the desired quality.

RQ1. Can self-managed social honeypots generate engagement on Instagram?
RQ2. How do the topic selection, post generation strategy, and engagement plan affect the success of a social honeypot?
RQ3. How much effort (computation and costs) is required to build an effective social honeypot?

The remainder of the section describes the strategies we adopt in our investigation, along with technical implementation details.

3.2 Topic Selection

Building a honeypot begins with selecting the topic of its posts. Such a choice will impact the type of users we will attract. The topic's nature might vary, from hobbies and passions like sports and music to sensitive issues like political views and conspiracies. As an example, if we wish to promote a new product of a particular brand, the topic might be the type of product we intend to promote. Alternatively, if we intend to develop a tool for spam detection, we should choose a topic that is interesting to spammers. This will ensure that they will be attracted to the honeypot's content. We can even design honeypots with generic topics that can be used for marketing profiling or social studies. In conclusion, the topic should be chosen in accordance with the honeypot's ultimate purpose.

3.3 Post Generation Strategies

The generative process aims to create posts pertaining to the honeypot topic. A two-part artifact is produced: the *visual* component of the post (i.e., the image), and its *caption*. We propose four distinct methods to generate posts, each with its own characteristics and algorithms. For ethical reasons, we excluded techniques that might violate the author's copyright (e.g., re-posting). However, unscrupulous honeypot creators could conveniently use these strategies. In this section, we provide the strategies high-level view to serve as a framework. For technical implementation details (e.g., the actual models we used), please refer to Appendix A. Since this stage involves deep generative models that might

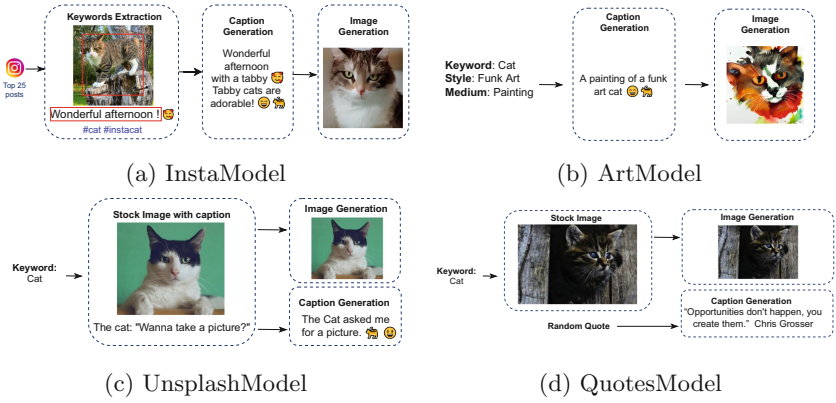

Fig. 2. Overview of Post Generation strategies.

produce artifacts affecting the post quality, the owner can approve a post or request a new one with negligible effort.

InstaModel. *InstaModel* is a generative schema that leverages machine learning techniques to generate both images and captions. Figure 2a shows its overview. The schema begins by retrieving one starting post among the 25 most popular IG posts for a popular hashtag related to the honeypot topic.[2] Next, the pipeline performs, in order, caption generation and image generation steps.

- *Caption Generation.* The algorithm uses an *Object Detector* tool[3] to extract the relevant elements of the starting post's image. In the absence of meaningful information (e.g., is a meme or unrelated to the topic)[4], we discard that image. When this occurs, the algorithm restarts and uses another sample from the top 25. If the image is kept, the algorithm uses the list of resulting elements (i.e., keywords) to generate a sentence, leveraging a *keyword-to-text* algorithm. Note that we discard from the keywords list those elements with very low probability. The output of the *keyword-to-text* phase (i.e., the new caption) is further refined to align with IG captions, for example, by adding emojis and hashtags, as presented in §3.4.
- *Image Generation.* The caption generated in the previous step serves as input to produce the post image. To achieve this goal, we use *text-to-image* models, i.e., algorithms that produce more images from a single input. An operator

[2] Starting from the main topic hashtags (i.e., #cat, #food, #car), we daily create the set of hashtags contained in the top 25 posts, from which we draw the hashtag to retrieve the starting post.

[3] Object detectors are Computer Vision-based tools that identify objects composing a given scene. Each object is accompanied by a probability score.

[4] We discard those images that do not contain at least a topic-related element with a high probability.

would choose the most appropriate option or a random option in such a case. We remark that *InstaModel* severely adopts generative models. Indeed, we used state-of-the-art computer vision, NLP, and image generation models for object detection, text generation, and image generation, respectively.

ArtModel. *ArtModel* leverages the ability of novel *text-to-image* generative models (e.g., DALL-E) to interpret artistic keywords as inputs. Figure 2b shows the overview of the model. Similarly to InstaModel, the process starts by generating a caption, and, subsequently, the image.

- *Caption Generation.* Differently from *InstaModel*, the input to generate the caption does not come from other IG posts. Instead, we randomly select the target keyword (e.g., cat), the artistic style of the picture (e.g., Picasso, impressionism), and a medium (e.g., painting, sketch). We create a single sentence by filling pre-defined templates with such three keywords, and add emojis and hashtags as for *InstaModel*.
- *Image Generation.* Similar to *InstaModel*, the caption (without emojis and hashtags) serves as input for a *text-to-image* model, which generates the final image.

UnsplashModel. This algorithm employs DL models only to generate the caption. In opposition to *InstaModel* and *ArtModel*, *UnsplashModel* starts from the image generation, and then generates the caption (Fig. 2c).

- *Image Generation.* The image is randomly selected by a stock images website – in this case, Unsplash[5]. The search is based on a randomly selected keyword that reflects the target topic, from a list defined by the owner.
- *Caption Generation.* Unsplash images are usually accompanied by captions free of license. We further refine the caption with a *rephrase* model, and add emojis and hashtags as for the previous models.

QuotesModel. Last, we present *QuotesModel*, a variant of *UnsplashModel*, presented in Fig. 2d. The objective of this strategy is to determine whether AI-based techniques are necessary to generate attractive IG posts. Therefore, this model does not involve the use of artificial intelligence to create captions and images. In addition, using quotes to caption photos is a diffused strategy [22].

- *Image Generation.* The image generation process is the same as *UnsplashModel*, involving stock images.
- *Caption Generation.* Captions are randomly selected by popular quotes from famous people (e.g., 'Stay hungry, stay foolish' – Steve Jobs). Quotes are retrieved from a pool with 1665 quotes [54].

[5] https://unsplash.com/.

3.4 Engagement Plans

Lastly, the engagement plan defines how the social honeypot interacts with the rest of the social network (e.g., other users or pages). We defined three plans, varying in effort required to maintain interactions, and whether paid strategies are involved:

- *PLAN 0*: low interactions and no paid strategies;
- *PLAN 1*: high interactions and no paid strategies;
- *PLAN 2*: high interactions and paid strategies.

PLAN 0. The plan does not involve automatic interactions with the rest of the social network. At most, the owner replies to comments left under the honeypot's posts. The plan uses the well-known *Call To Actions* (CTA) [39] in the posts. Such a strategy consists in creating captions that stimulate users' engagement (e.g., liking, commenting, sharing the post). Examples are captions containing simple questions (e.g., 'How was your day?'), polls and quizzes (e.g., 'What should I post next?), or exhorting users to share their opinions (e.g., 'What do you think about it?'). Following the caption best strategies for IG posts [46], we added 15 random hashtags related to our topic, 8 with broad coverage and 7 with medium-low coverage. More details about the hashtags selections in Appendix A. In this plan, paid strategies are not involved.

PLAN 1. The plan is a variant of *PLAN 0* with explicit social networking interactions. We call these actions *spamming*. The spamming consists of automatically leaving likes and comments on the top 25 posts related to the topic (as described in *InstaModel*). Comments resemble legit users (e.g., 'So pretty!') and not spammers (e.g., 'Follow my page!'), and were randomly picked from a list we manually created by observing comments usually left under popular posts. The goal of such activities is to generate engagement with the owner of popular posts, hoping to redirect this stream to the honeypot. When a user follows us, we follow back with a probability of 0.5, increasing the page's number of followings, resembling a legit page. During our experiments, we also adopted a more aggressive (and effective) spamming strategy called *Follow & Unfollow* (F&U) [13], consisting in randomly following users, often causing a follow back, and then remove the following after a couple of days. To not be labeled as spammers, we constantly respected the balance # following $<$ #followers. In this plan, paid strategies are not involved.

PLAN 2. This plan increments *PLAN 1* with two paid strategies.

Buying followers. When we create a honeypot, we buy N followers. In theory, highly followed pages might encourage users to engage more, and gain visibility from IG algorithm [65]. Therefore, we aim to understand if an initial boost of followers can advantage honeypots. Such followers will be discarded during our analyses. We set $N = 100$, and we buy passive followers only.[6]

[6] Passive followers only follow the page, but they do not engage further.

Content sponsoring. IG allows posts' sponsoring for a certain amount of time. The target population can be automatically defined by IG, or chosen by the owner w.r.t. age, location, and interests. Since we are interested in studying the population attracted by our content, rather than attracting a specific category of users, we let IG decide our audience, directly exploiting its algorithms to make our honeypots successful.

4 Implementation

4.1 Topic Selection

We investigate the honeypots' effectiveness over three distinct topics: *food*, *cat*, and *car*. We selected such topics to account for different audience sizes, measured by coverage levels. Coverage is a metric that counts the total number of posts per hashtag or, in other words, the total number of posts that contain that hashtag in their captions. This information is available on IG by just browsing the hashtag. More in detail, we selected: **Food** (high coverage, #food counts 493 million posts), **Cat** (medium coverage, #cat counts 270 million posts), and **Car** (low coverage, #car counts 93 million posts). We chose these topics, and not more sensitive ones, mainly for ethical reasons. Indeed, we did not want to boost phenomena like misinformation or conspiracies through our posts, nor identify people involved in these themes. However, we designed our methodology to be as general as possible, and adaptable to any topic with little effort.

4.2 Testbed

We deployed 21 honeypots on Instagram, seven for each selected topic (i.e., food, cat, and car), that we maintained for a total of nine weeks. Within each topic, we adopt all post generation strategies and engagement plans. For the post generation strategies, three honeypots use both InstaModel and ArtModel, three honeypots use UnsplashModel and QuotesModel, and one honeypot combines the four. Such division is based on the image generation strategy, i.e., if images are generated with or without Deep Learning algorithms. All posts were manually checked before uploading them on Instagram to prevent the diffusion of harmful or low-quality content. This was especially necessary for AI-generated content, whose low quality might have invalidated a fair comparison with non-AI content.[7] Similarly, for the engagement plan, two honeypots adopt PLAN 0, two PLAN 1, and three PLAN 2. Table 1 summarizes the 21 honeypots settings. Given the nature of our post generation strategies and engagement plans, we set as baselines the honeypots involving *UnsplashModel + QuotesModel* as generation strategy and *PLAN 0* as engagement plan (h1, h8, h15). Indeed, these

[7] The effort for the honeypot manager is limited to a quick approval, which could not be necessary with more advanced state-of-the-art models, e.g., DALL-E 2 [1] or ChatGPT [52].

honeypots are the simplest ones, requiring almost no effort from the owner. Setting baselines is useful to appreciate the results of more complex methods, given that there are currently no baselines in the literature.

By following the most common guidelines [48,63], each honeypot was designed to publish two posts per day, with at least 8 h apart from each other.

During the nine weeks of experiments, we varied PLAN 1 and PLAN 2. In particular, we started PLAN 1 with spamming only, and PLAN 2 with buying followers. During the last week, both plans adopted more aggressive strategies, specifically, PLAN 1 applied F&U techniques, while PLAN 2 sponsored the two most-popular honeypot posts for one week, paying €2/day for each post. For our analyses, we collected the following information:

- Total number of followers per day;
- Total number of likes per post;
- Total number of comments per post.

Moreover, IG API provided the gender, age, and geographical locations of the audience when applicable, as explained in §6.3.

Table 1. Honeypots deployed.

ID	Post Generation Strategy	Engagement Plan
	food	
h1 (baseline)	UnsplashModel + QuotesModel	PLAN 0
h2	UnsplashModel + QuotesModel	PLAN 1
h3	UnsplashModel + QuotesModel	PLAN 2
h4	InstaModel + ArtModel	PLAN 0
h5	InstaModel + ArtModel	PLAN 1
h6	InstaModel + ArtModel	PLAN 2
h7	All Models	PLAN 2
	cat	
h8 (baseline)	UnsplashModel + QuotesModel	PLAN 0
h9	UnsplashModel + QuotesModel	PLAN 1
h10	UnsplashModel + QuotesModel	PLAN 2
h11	InstaModel + ArtModel	PLAN 0
h12	InstaModel + ArtModel	PLAN 1
h13	InstaModel + ArtModel	PLAN 2
h14	All Models	PLAN 2
	car	
h15 (baseline)	UnsplashModel + QuotesModel	PLAN 0
h16	UnsplashModel + QuotesModel	PLAN 1
h17	UnsplashModel + QuotesModel	PLAN 2
h18	InstaModel + ArtModel	PLAN 0
h19	InstaModel + ArtModel	PLAN 1
h20	InstaModel + ArtModel	PLAN 2
h21	All Models	PLAN 2

Implementation Models. In §3 we presented a general framework to create social honeypots. In our implementations, we employed deep learning state-of-the-art models in several steps. To extract keywords in *InstaModel* we adopted InceptionV3 [67] as object detector, pre-trained on ImageNet [17] with 1000 classes. From the original caption, we extracted nouns and adjectives through NLTK python library[8]. As *keyword-to-text* algorithm, we adopted Keytotext [21] based on T5 model [55]; while for *text-to-image* processes we opted for Dall-E Mini [14]. Finally, in *UnsplashModel*, the rephrase task was performed using the Pegasus model [77].

5 Honeypots Evaluation

5.1 Overall Performance

The first research question *RQ1* is whether social honeypots are capable of generating engagement. After nine weeks of execution, our 21 social honeypots gained: 753 followers (avg 35.86 per honeypot), 5387 comments (avg 2.01 per post), and 15730 likes (avg 5.94 per post). More in detail, Table 2 (left side) shows the overall engagement performance at the varying of our three variables, i.e., topic, generation strategy, and engagement plan. The reader might notice that not only our honeypots *can* generate engagement, answering positively to the *RQ1*, but that also topic, generation strategy, and engagement plan have different impacts to the outcomes. For instance, *cat* honeypots tend to have higher followers and likes, while *car* ones generate more comments. Similarly, *non-AI* generation methods tend to have higher likes, as well as *PLAN 1*. We investigate the effect of different combinations later in this section.

5.2 Honeypot Trends Analysis

Social honeypots can generate engagement, but we are further interested in understanding trends of such performance: *is honeypots' engagement growing over time?* A honeypot with a positive trend will likely result in a higher future attraction. On the opposite, a stationary trend implies limited opportunities to improve.

The qualitative analysis reported in Fig. 3 motivates the trend investigation. The figure presents the average number of Likes per post gained by our honeypots over time, grouped by engagement plan. In general, PLAN 1 honeypots tend to attract more likes as they grow, followed by PLAN 2 and PLAN 0, in order. In particular, a constantly increasing number of likes is shown by honeypots with PLAN 1, especially for food-related pages: starting from an average of ∼5 likes per post (week 1st) to ∼12.5 likes per post (week 9th). We evaluate the presence of stationary trends by adopting the *Augmented Dickey-Fuller test* (ADF) [51]. In this statistical test, the null hypothesis H_0 suggests, if rejected, the presence of a non-stationary time series. On the opposite, the alternative hypothesis H_1

[8] https://www.nltk.org/.

Table 2. Honeypots overall performance. On the left side, we report the average (and std) engagement generated by the honeypots. On the right, we report the number of honeypots with a non-stationary trend. The results are reported based on the topic, generation strategy, and engagement plan.

	Average Engagement			**Engagement Trend**		
	#Followers	#Comments	#Likes	#Followers	#Comments	#Likes
			topic			
food	38.5±33.7	216.4±18.5	698.4±139.7	6/7	3/7	7/7
cat	**47.4**±17.5	182.1±23.5	**923.1**±214.8	6/7	2/7	4/7
car	21.9±9.7	**371.0**±26.2	625.6±96.6	7/7	3/7	6/7
			generation strategy			
AI	37.9±30.9	248.4±94.6	654.2±138.3	7/9	4/9	6/9
non-AI	32.7±21.3	**264.2**±90.6	**842.5**±235.2	9/9	3/9	8/9
Mixed	**39.3**±7.9	257.7±80.0	753.0±125.9	3/3	1/3	3/3
			engagement plan			
PLAN 0	11.5±8.4	**266.0**±105.8	641.3±210.7	4/6	4/6	5/6
PLAN 1	**60.0**±25.2	254.2±94.3	**835.2**±210.7	6/6	2/6	4/6
PLAN 2	36.0±14.0	251.8±79.1	763.4±206.1	9/9	2/9	8/9

suggests, if rejected, the presence of a stationary time series. We conducted the statistical test for each honeypot and the three engagement metrics: #Followers, #Likes, and #Comments. A p-value > 0.05 is used as a threshold to understand if we fail to reject H_0. Table 2 (right side) reports the result of the analysis. The number of Followers and Likes is non-stationary in 19 and 17 cases out of 21, respectively. Conversely, the number of comments per post is stationary in most of the honeypots. This outcome suggests that engagement in terms of likes and followers varies over time (positively or negatively), while the number of comments is generally constant. As shown in Fig. 3, and given the final number of followers higher than 0 (i.e., at creation time), we can conclude that our honeypots present, in general, a growing engagement trend.

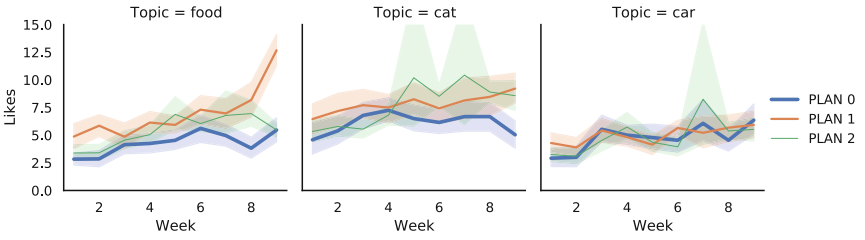

Fig. 3. Likes trend of our honeypots grouped by engagement plan.

5.3 The Impact of Honeypots Configuration

We now investigate whether the three variables (i.e., topic, generation strategy, and engagement plan) have a statistical impact on the success of the honeypots, answering *RQ2* and *RQ3*. Given the stationary trend of comments, we focus solely on likes per post and followers per honeypot.

Likes. Figure 4 depicts the distribution of honeypots Likes at the varying of the topic, generation strategy, and engagement plan. In general, there is a difference when the three variables are combined. For example, on average, honeypots belonging to cats, with non-AI generative models, and with PLAN1 or PLAN2 have higher values than the rest of the honeypots. Moreover, in general, honeypots adopting PLAN1 have higher results.

To better understand the different impacts the three variables have on Likes, we conducted a three-way ANOVA. We found that both topic, engagement plan, and generation strategy are significantly (p-value < 0.001) influencing the Likes. Furthermore, we found significance even in the combination of topic and engagement plan (p-value < 0.001), but not in the other combinations. This result confirms the qualitative outcomes we have presented so far. We conclude the analysis by understanding which topic, generation strategy, and engagement plan are more effective. To this aim, we performed Tukey's HSD (honestly significant difference) test with significance level $\alpha = 5\%$. Among the three topics, *cat* is significantly more influential than both *food* and *car* (p-value $= 0.001$). Regarding the generation strategies, non-AI-based models (i.e., UnsplashModel and InstaModel) outperform AI-based ones. Last, PLAN1 and PLAN2 outperform PLAN0 (p-value $= 0.001$), while the two plans do not show statistical differences between them.

Followers. Tukey's HSD test revealed statistical differences in the number of followers as well. For the analysis, we use the number of followers of each honeypot at the end of the 9th week. We found that *cat* statistically differ from *car* (p-value < 0.01), while there are no significant differences between *cat* and *food*, or *food* and *car*. Regarding the generation strategy, we found no statistical difference among the groups. Finally, all three engagement plans have a significant impact on the number of followers (p-value $= 0.001$), where PLAN 1 $>$ PLAN 2 $>$ PLAN 0.

Aggressive Engagement Plans. We recall that honeypots deployed with PLAN 1 and PLAN 2 adopted more aggressive engagement strategies on week 9th: *Follow & Unfollow* for PLAN 1, and *Content Sponsoring* for PLAN 2. Thus, we investigated whether aggressive plans result in more engagement in terms of comments, likes, and followers. The analysis is performed with Tukey's HSD (honestly significant difference) test with significance level $\alpha = 5\%$. We found no statistical difference in comments in PLAN 1 and PLAN 2. On the opposite, the average number of likes per post shows a statistically significant improvement

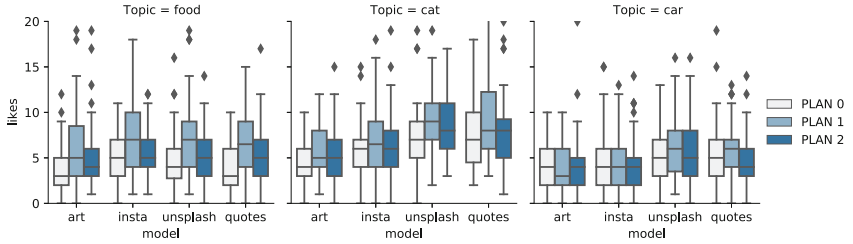

Fig. 4. Distribution of likes at the varying of topic, model generation strategy, and engagement plan.

in PLAN1 (p-value = 0.01): on average, 7.44 and 9.17 likes per post in weeks 8th and 9th, respectively. No statistical difference is found for PLAN 2; indeed, only the sponsored content benefited (i.e., a few posts).[9] Last, we analyze the difference between the total amount of followers at the end of weeks 8th and 9th. PLAN 1 honeypots #Followers moved, on average, from 45.7 ± 19.1 of week 8th, to 60.7 ± 26.2 of week 9th, with no statistical difference. PLAN 2 honeypots #Followers moved, on average, from 22.3 ± 11.6 of week 8th, to 30.7 ± 13.9 of week 9th. The difference is statistically supported (p-value < 0.05).

5.4 Baseline Comparison

Social honeypots are effective, depending on topics, generation strategies, and engagement plans. Since we are the first, to the best of our knowledge, to examine how to attract *people* using social honeypots (not bots or spammers), there are no state-of-the-art baselines to compare with. Therefore, we compare our methodology with (i) our proposed non-AI generative models with a PLAN 0 engagement strategy (baseline) and (ii) real Instagram pages trends.

Baseline. This represents the most simplistic method someone might adopt: adding stock images, with random quotes, without caring about the engagement with the rest of the social network. From §5.3, we statistically showed that the definition of engaging plans is essential to boost engagement in social honeypots. We remark on this concept with Figs. 5 and 6 that show the comparison among the baselines and PLAN 1 social honeypot – which are the most effective ones – in terms of likes and followers over the 9 weeks: in terms of AI and Non-AI strategies, our advanced honeypots outperform in 3 out of 6 cases and 6 out of 6 cases the baselines for likes and followers, respectively. Such results confirm the remarkable performance of our proposed framework. Our strategies might perform worse than the baselines (regarding likes) when the image quality is unsatisfactory. Indeed, as demonstrated in our prior work [68], likes on IG are usually an immediate positive reaction to the post's image. Since Unsplash images are

[9] All sponsored content belongs to weeks before the 9th.

usually high-quality and attractive, they might have been more appealing than AI-generated images in these cases.

Although comparing our approach with other social honeypots [42,75,79] carries some inherent bias (the purpose and social networks are completely different), we still find our approach aligned with (or even superior than) the literature. Lee et al. [42] gained in seven months through 60 honeypots a total of ∼36000 interactions (e.g., follow, retweet, likes), which is approximately 21.5 interactions per honeypot/week. Our honeypots reached a total of 21870 interactions, which is approximately 115.7 interactions per honeypot/week, i.e., more than five times higher. Yang et al. [75] lured 1512 accounts in five months using 96 honeypots, i.e., 0.788 accounts per honeypot/week. We collected 753 followers, which is 3.98 accounts per honeypot/week, i.e., five times higher. Last, Zhang et al. [79] carefully selected and harnessed 1000 popular Twitter accounts (which they called pseudo-honeypots) for three weeks to analyze spammers. Giving these accounts were already heavily integrated into the social network, they reached over 476000 users, which is around 159 accounts per (pseudo-)honeypot per week. We remind that the purpose of these comparisons is to give an idea of the effectiveness of other social honeypots rather than to provide meaningful conclusions.

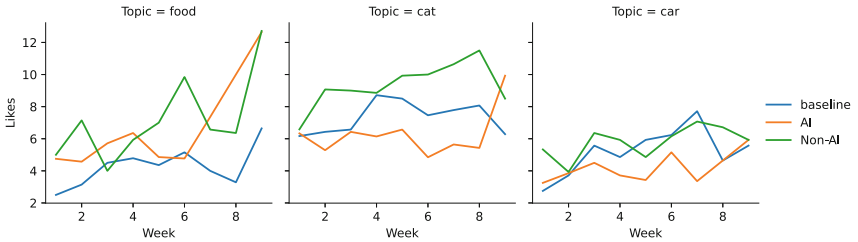

Fig. 5. Baseline comparison (average likes) with PLAN1 social honeypots.

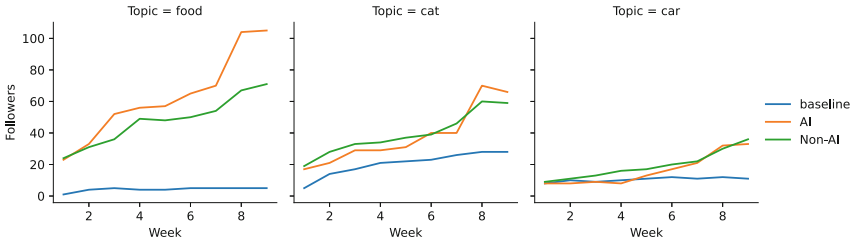

Fig. 6. Baseline comparison (followers) with PLAN1 social honeypots.

Instagram Pages. We now compare our PLAN 1 social honeypots with real IG public accounts. Accordingly, we analyzed the first nine weeks of activities on popular IG pages related to food, cat, and cars. We selected nine popular IG pages for each topic, 3 with $\sim 10K$ followers, 3 with $\sim 100K$ followers, and 3 with more than a million followers. We collected the number of comments and likes for each post published during this period. Due to IG limitations, we could access only information at the time of collection, implying that posts might be a few years old. Monitoring new pages would be meaningless since we do not know a priori whether they will become popular.

We noticed that it is impossible to compare such baselines with our social honeypots because, generally, the considered IG pages contain posts with hundreds of likes and comments even in their first week of activity. For instance, $+1M$ pages' first posts reached more than 2000 likes. Possible explanations behind this phenomenon are: (i) the considered 18 pages were already popular before their creation (e.g., on a different or older OSN like Facebook); (ii) the considered 18 pages massively sponsored all their content; (iii) we are facing the *earlybird bias*, where older posts contain not just engagement from the first nine weeks, but also engagement from later periods, even years.[10] To further explain this phenomenon, we contacted such IG pages (we extended our survey to 36 pages). Questions focused on the first weeks of activity.[11] Unfortunately, up to the submission date, none of the contacted pages replied.

Although there is no evidence in the literature on how long it takes to make an Instagram page famous, most sources consider the initial growth (from 0 to 1000 followers) to be the most challenging part [5,27], with an overall monthly growth rate of about 2% [44]. Furthermore, success requires lots of dedication to follow best practices consistently [47], which is extraordinarily time-consuming and far from trivial. Being in line with these trends in a fully automated and effortless manner is already an impressive achievement. Our work can serve as a baseline and inspiration for future work.

6 Social Analyses

6.1 Comments Analysis

An interesting (and unexpected) result is that, without the premeditated intention of building spammer detectors, most of the comments we received came from spammers. To estimate the total number of spam comments, we first manually identified patterns used by spammers on our honeypots (e.g., expressions like "send pic" or "DM us"). Afterward, using a pattern-matching approach, we found that 95.33% of the comments we received on our social honeypots came indeed from spammers. All spammers' accounts shared similar behavior in

[10] Earlybird bias appears in other social contexts like online reviews [43].

[11] For instance, we asked whether the page resulted from an already existing page (on IG or other platforms), or the strategies they adopted to manage the pages (e.g., spam, sponsoring).

commenting: (i) there was always a mention '@' to other accounts, and (ii) they commented almost immediately after the post creation. Such considerations suggest these accounts are bots that target many recent posts, perhaps searching by specific hashtags. Such findings indicate that fresh pages could be a powerful tool to detect spammers with *minimal* effort. We also highlight that spam comments are a well-known issue that affects the majority of IG pages [40] and is not limited to our honeypot pages. Therefore, we argue that creating pages that do not attract spammers is nearly impossible. Nevertheless, IG itself is employing and improving automatic screening mechanisms [31,32] to limit such behavior. When such mechanisms are enhanced, our honeypots will become more accurate.

6.2 Followers Analysis

As most of our comments were spam, we investigated whether followers were the same. We manually inspected the followers of our most followed social honeypot for each topic, identifying three categories of followers:

- *Real people:* users that publish general-topic posts, with less than 1000 followers[12], and real profile pictures;
- *Pages and Influencers:* users that publish topic-specific posts (e.g., our honeypots) or with more than 1000 followers;
- *Bots:* users whose characteristics resemble a bot, following well-known guidelines [3], e.g., fake or absent profile picture, random username, highly imbalanced follower/following count, zero or few (< 5) posts.

From Table 3, we notice the three honeypots have different audiences. The *food* honeypot obtained the most real followers, *car* reached more bots, and *cat*, was followed mainly by pages. These results confirmed that (i) our honeypots can reach real people, (ii) the audience category depends on the topic, and (iii) spammers' threat is limited to comments. On an interesting note, most pages following our *cat* honeypot were cat-related pages.

Table 3. Percentage of real people, pages, and bots for the best social honeypot in each topic.

	Real People	**Pages**	**Bots**
Food	**48,08**%	37,50%	14,42%
Cat	10,61%	**72,72**%	16,67%
Car	30,30%	21,21%	**48,49**%

6.3 Reached Audience

We conclude the experimental results with a detailed analysis of the audience our honeypots reached. In particular, we performed two distinct analyses: (i)

[12] After 1000 followers, users are considered nano influencers [53].

Honeypot reached audience, and (ii) *Sponsored posts audience*, i.e., IG features available for honeypots with 100 followers and sponsored content, respectively. After nine weeks of computation, one honeypot satisfies the requirement of 100 followers (honeypot ID: h9). About the sponsored content, we obtained information about 9 posts (one per honeypot belonging to PLAN 2).

Honeypot Audience. The honeypot h9 (topic: food, generation strategy: AI, and engagement plan: PLAN 1) gained 103 followers: the majority is distributed over the age range $[25-34]$ with 32% (equally distributed among men and women), $[35, 44]$ with 10% of women and 27% of men. Most followers came from India (11.7%), Bangladesh (10.7%), and Japan (9.7%).

Sponsored Posts Audience. For this analysis, we recall that we set our sponsoring strategy leveraging the automatic algorithm provided by IG. Overall, sponsored posts achieved great success in terms of generated engagement. On average, food posts reached 30.6, 116, and 60.6 likes for food, cat, and car posts, respectively. These numbers are strongly above the average likes per post 5.9. IG offers an analytic tool to inspect the reached audience; this feature perfectly fits in the scope of social honeypots, since it allows finding insights about the attracted audience. For each post, the following information is available: quantitative information (i.e., reached people, likes, comments, sharing, saved), and demographic distribution in percentage (gender, age, location). The detailed report is available in Appendix B. We observed interesting trends:

- *food* audience: the gender is almost balanced (female audience slightly more attracted), and the predominant age range is 18–34. Top locations: Campania, Lombardia, and Puglia.[13]
- *cat* audience: the gender distribution is toward the female sex, and the predominant age range is 18-34. Top locations: Emilia Romagna, Lombardia, Piemonte.
- *car* audience: the gender is strongly distributed toward the male sex, and the predominant age range is 18-24. Top locations: Lazio, Lombardia.

To conclude, with minimal effort (i.e., €2/day per post), an owner can get useful information, e.g., to use in marketing strategies..

7 Toward a Real Implementation

So far, we have demonstrated our social honeypots can attract real people in a fully automated way. With little effort, they can already be deployed for an array of situations. In this section, we first reason about the use cases of our approach, highlighting both positive and negative outcomes. Then, we present the current challenges and limitations of implementing this work in real scenarios.

[13] IG automatic algorithm maximized the audience toward authors country, i.e., Italy, reporting Italian regions.

7.1 Use Cases

Our work aims to show the lights and shadows of social networks such as Instagram. People can easily deploy automated social honeypots that can attract engagement from hundreds or even thousands of users. Upon on that, analyses on these (unaware) users can be conducted. As cyber security practitioners, we know that this technology might be exploited not only for benign purposes, but also to harm users [74]. Therefore, this work contributes to the discussion about the responsible use of online social networks, in an era when technologies like artificial intelligence are transforming cyber security. We list in this section possible social honeypot applications.

Marketing. The first natural adoption of our proposed social honeypots is for marketing purposes. Suppose someone is interested in understanding "who is the average person that loves a specific theme", where themes might be music, art, puppies, or food. With a deployed social honeypot, the owner can then analyze the reached audience by using the tools offered by IG itself (as we ethically did in this paper) or by further gathering (potentially private) information on the users' profile pages [70].

Phishing and Scam. Similarly to marketing, social honeypots can be used by adversaries to conduct phishing and scam campaigns on IG users. For instance, the social honeypot might focus on cryptocurrency trading: once identified potential victims, attackers can target them aiming to obtain sensitive information (e.g., credentials), or to lure them into fraudulent activities such as investment scams, rug pulls, Ponzi schemes, or phishing.

Spammer Identification. Social honeypots can also be created to imitate social network users, by posting content and interacting with other users. As we noticed in our experiments, they can attract spammers. Therefore, our proposed framework can be adopted by researchers to spot and study new types of spamming activities in social networks.

Monitoring of Sensible Themes. An interesting application of social honeypots is to identify users related to sensible themes and monitor their activities (within the honeypot). Examples of such themes are fake news and extremism [57]. Researchers or authorities might leverage social honeypots to identify users that actively follow and participate in such themes, and then carefully examine their activity. For instance, honeypot owners can monitor how people respond to specific news or interact inside the honeypot.

7.2 Challenges and Limitations

The first challenge we faced in our work is the massive presence of spammers on IG. Most of them are automated accounts that react to particular hashtags and

comments under a post for advertisement or scamming purposes [38,78]. This factor can inevitably limit our approach when we aim to gather only real people. As a countermeasure, honeypots should include a spam detector (e.g. [26,78]) to automatically remove spammers. On the contrary, this approach could be useful directly to reduce the spamming phenomenon. Many pages can be created with the purpose of attracting spammers and reporting them to IG for removal.

The second challenge we encountered is the lack of similar works in the literature. Because of this, we have no existing baselines to compare with, and it could be difficult to understand whether our approach is truly successful. However, in nine weeks, we obtained more than 15k likes and gathered $\sim$ 750 followers in total, which is not trivial as discussed in §5.4. Our most complex methods surpassed the simplest strategies we identify, which can serve as a baseline and source of inspiration for future works.

Among the limitations, we inspected only generic (and ethical) topics. A comprehensive study in this direction would give much more value to our work, especially dealing with delicate topics (e.g., conspiracies, fake news). Moreover, our approach is currently deployable on IG, but would be hard to transfer to other platforms. Even if this can be perceived as a limitation, it would be naive to consider all social media to be the same. Indeed, each of them has its own content, purpose, and audience. Developing social honeypots for multiple platforms can be extremely challenging, which is a good focus for future research. Last, there was no clear connection between the posts of our honeypots. When dealing with specific topics, it might be necessary to integrate more cohesive content.

8 Conclusions

The primary goal of this work was to first understand the feasibility of deploying self-managed Instagram Social Honeypots, and we demonstrated that *it is possible* in §5.1. Moreover, from the results obtained in our analyses we can derive the following outcomes and guidelines:

1. *Topics* plays an important role in the success of the honeypot.
2. *Generation strategies* does not require complex DL-based models, but simple solutions such as stock images are enough. Similarly, we saw that posts containing random quotes as captions are as effective as captions describing the content;
3. *Engagement plan* is essential. We demonstrated that a naive engagement strategy (PLAN 0) results in a low volume of likes and followers. Moreover, the engagement plan without costly operations (PLAN 1) works as well as plans involving followers acquisition and content sponsoring;
4. *Sponsored content* is a useful resource to preliminary assess the audience related to a specific topic;
5. Social honeypots not only attract *legitimate* users, but also *spammers*. As a result, they can be adopted even for cybersecurity purposes. Future implementation of social honeypots might include automatic tools to distinguish engagement generated by legitimate and illegitimate users.

In conclusion, we believe that our work can represent an important milestone for future researchers to easily deploy and collect social network users' preferences. New research directions might include not only general topics like cats and food, but more sensitive themes like fake news, or hate speech. In the future, we expect generative models to be always more efficient (e.g., DALL-E 2 [1] or ChatGPT [52]), thus increasing the reliability of our approach (or perhaps making it even more dangerous).

Ethical Considerations

Our institutions do not require any formal IRB approval to carry out the experiments described herein. Nonetheless, we designed our experiments to harm OSN users as less as possible, adhering to guidelines for building Ethical Social Honeypots [20], based on the Menlo report [6]. Moreover, we dealt with topics (cars, cats, food) that should not hurt any person's sensibility. In our work, we faced two ethical challenges: data collection and the use of deception. Similar to previous works [15,42,75], we collected only openly available data (provided by Instagram), thus no personal information was extracted, and only aggregated statistics were analyzed. Moreover, all information is kept confidential and no-redistributed. Upon completion of this study, all collected data will be deleted. This approach complies with the GDPR. To understand the honeypot's effectiveness, similar to previous works, we could not inform users interacting with them about the study, to limit the Hawthorne effect [24]. However, we will inform the deceived people at the end of the study, as suggested by the Menlo report.

A Implementation Details

A.1 Models

In this appendix we will describe how InstaModel, ArtModel, UnsplashModel and QuotesModel were implemented. All of them have different characteristics but, at the same time, share some common functionalities that will be explained before of the actual implementation of the four models.

Shared functionalities. One of the shared functionalities is adding emojis to the generated text. This is done with a python script which scans the generated caption trying to find out if there are words that can be translated with the corresponding emoji. To make this script more effective, it looks also for synonyms of nouns and adjectives found in the text to figure out if any of them can be correlated to a particular emoji. As last operation, the script chooses randomly, from a pool of emojis representing the "joy" sentiment, one emoji for each sentence that will be append at the end of each of them.

CTA are simple texts that may encourage a user to do actions. These CTA are sampled randomly from a manually compiled list and then added at the end of the generated caption.

The last shared feature is the selection of hashtags. As said before, through the Instagram Graph API we are able to get the first 25 posts for a specific hashtag and from them we extracted all the hashtags contained in the caption. Thus we compiled an hashtag list for each of the three topic sorted from the most used to the least used. Instagram allows to insert at most 30 hashtags in each posts but we think that this number is too high with respect to the normal user's behavior. For this reason, we decided to choose 15 hashtags that are chosen with this criteria: 8 hashtags are sampled randomly from the first half of the list in the csv file, giving more weight to the top ones, while the other 7 are sampled randomly from the second half of the list, giving more weight to the bottom part of the list. The intuition is that we are selecting the most popular hashtags together with more specific hashtags.

InstaModel. Starting from the caption generation, InstaModel uses the Instagram Graph API to retrieve the top 25 posts for a specific hashtag. In practice, the chosen hashtag will be the topic on which the corresponding honeypot is based. Once we have all the 25 posts, they are checked to save only those that have an English caption before being passed to the object detector block. The object detector is implemented by using the InceptionV3 model for object detection tasks. InceptionV3 detects, in the original image, the object classes with the corresponding accuracy and if the first's class score is not greater than or equal to 0.25, the post will be discarded. Otherwise, the other classes are checked as well and only if their scores are greater than 0.05 will be considered as keywords for the next step. Regarding the original caption, nouns and adjectives are extracted by using nltk python library. Notice that words such as "DM" or "credits" and adjectives such as "double" or similar, are not considered. This is because they usually belong to part of the caption that is not useful for this process.

Keyword2text[14] is the NLP model that transforms a list of keywords in a preliminary sentence. This preliminary sentence is then used by OPT model to generate the complete text. Considering the computational resources available to us, the model used is OPT with 1.3 billion parameters. We suggest to save the text generated by OPT in a file text because it will be used subsequently to generate the corresponding image. Once we have the complete generated text, emojis are added together with a CTA sentence that is standard in any post. The last step for caption generation is to append hashtags: they will be chosen by sampling from the corresponding csv file with the reasoning mentioned above.

The last step of InstaModel is image generation and for this purpose Dall-E Mini ([14]) is used. The prompt will be the text generated after the OPT stage, the one that has been save separately. It is relevant to highlight that the process with Dall-E Mini is not completely automatic and there should be a person that choose the most suitable image for the giving caption.

ArtModel. ArtModel starts from a prompt generated with a python script and uses Dall-E mini, like InstaModel, to generate the corresponding image. The style

[14] https://huggingface.co/gagan3012/k2t.

and the medium are chosen randomly from two lists. Example of styles can be "cyberpunk", "psychedelic", "realistic" or "abstract" while examples of medium are "painting", "drawing", "sketch" or "graffiti". The topic of the honeypot is used as subject of the artistic picture generated by Dall-E Mini. Once the image is generated, the prompt, added of emojis, CTA and the corresponding hashtags, will be used as Instagram caption.

UnsplashModel. UnslashModel does not generate images but uses stock images retrieved from the Unsplash websites. Unsplash has been chosen not only because it gives the opportunity to find images together with the relative captions, but also because it offers API for developers that can be used easily. To avoid reusing the same images more than once, each image's id is saved in a text file which will be checked at each iteration. For the caption generation, the original caption is processed by Pegasus model ([77]) which is an NLP model quite good in the rephrase task. As always, emojis, CTA and hashtags are added to the final result.

QuotesModel. QuotesModel makes use of Pixabay[15] stock images website to avoid reusing Unsplash even for this model. Also in this case, we use the topic of the specific honeypot as query tag. As for UnsplashModel, to avoid reusing the same image for different posts, once we have downloaded the image, its id is saved in a text file which will be checked every time needed. For the caption generation, a quote is sampled randomly from a citation dataset [22]. In this case, the model does not add emojis to the text because we think that the quote, by itself, can be a valid Instagram caption. On the contrary, as always, CTA and hashtags are added to the text.

A.2 Spamming

Honeypots with PLAN 1 or PLAN 2 engagement plans will automatically interact with the posts of other users. The idea is to retrieve the top 25 Instagram posts for the hashtag corresponding to the specific topic of the honeypot and like and comment each of them.

For the implementation we used Selenium which is a tool to automates browsers and it can be easily installed with pip command. Selenium requires a driver to interface with the chosen browser and in our case, since we chose Firefox, we have downloaded the geckodriver. The implementation consists of a python class which has three main methods: `login`, `like_post` and `comment_post`

The login method is invoked when the honeypot accesses to Instagram. The like_post method searches, in the DOM, for the button corresponding to the like action and then it clicks it. The comment_post method searches in the DOM for the corresponding comment button and then clicks it. Afterwards, it searches for the dedicated textarea and write a random sampled comment. Finally, it clicks the button to send the comment.

[15] https://pixabay.com/.

Table 4. Overview of the sponsored content attracted users

Overview									
honeypot	h3	h6	h7	h10	h13	h14	h17	h20	h21
topic	food	food	food	cat	cat	cat	car	car	car
gen. strat.	AI	NON AI	NON AI	AI	NON AI	NON AI	AI	NON AI	NON AI
audience	3126	3412	5337	3245	4597	2863	10698	6824	9633
likes	21	34	37	118	163	67	20	25	127
comments	1	3	7	3	8	1	3	11	3
saved	1	0	21	12	29	7	2	6	44
Gender Coverage [%]									
women	42.2	60.0	87.8	67.2	67.7	59.0	8.6	8.7	5.6
men	57.0	38.7	11.7	31.5	30.7	39.3	89.5	90.7	93.6
Age Coverage [%]									
13 – 17	0.1	0.1	0	0	0	0.1	0.2	0.1	0.1
18 – 24	39.1	37.7	35.9	20.8	33.8	38.6	64.3	45.7	52.5
25 – 34	29.8	12.9	36.0	21.2	25.2	15.2	12.7	31.8	26.8
35 – 44	14.5	11.6	14.3	15.6	13.0	12.4	6.5	10.8	9.4
45 – 54	9.0	18.3	8.2	18.7	14.0	13.7	8.1	5.1	6.1
55 – 64	4.7	12.9	3.8	15.8	9.3	12.4	5.0	3.6	3.0
65+	2.5	6.0	1.3	7.5	4.3	7.2	2.9	2.6	1.8
Geographic Coverage [%]									
Campania	14.7	11.3	9.1	N.A	N.A	8.7	7.8	8.7	N.A
Emilia-Romagna	N.A	N.A	N.A	9.7	8.7	9.2	N.A	8.6	9.2
Lazio	N.A	7.9	8.3	9.4	10.5	N.A	8.2	11.1	9.5
Lombardia	12.4	12.0	13.2	19.6	18.8	17.2	14.0	19.0	20.9
Piemonte	N.A	N.A	N.A	9.0	8.5	7.5	N.A	N.A	8.0
Puglia	12.5	10.9	8.9	N.A	N.A	N.A	8.9	N.A	N.A
Sicilia	9.0	10.0	9.2	N.A	N.A	N.A	10.4	N.A	N.A
Tuscany	N.A	N.A	N.A	7.2	N.A	N.A	N.A	N.A	N.A
Veneto	9.0	N.A	N.A	N.A	7.7	8.4	N.A	8.8	10.1

B Sponsored Content Analyses

We report in Table 4 the complete overview of audience attracted by our sponsored content. In particular, we report overall statistics in term of quantity (e.g., number of likes), and demographic information like gender, age, and location distribution.

References

1. Aditya, R., Prafulla, D., Alex, N., Casey, C., Mark, C.: https://openai.com/product/dall-e-2 (2022), Accessed Mar 2023
2. Ahmed, W., Vidal-Alaball, J., Downing, J., Seguí, F.L., et al.: Covid-19 and the 5g conspiracy theory: social network analysis of twitter data. J. Med. Internet Res. **22**(5), e19458 (2020)
3. Akyon, F.C., Kalfaoglu, M.E.: Instagram fake and automated account detection. In: 2019 Innovations in intelligent systems and applications conference (ASYU). pp. 1–7. IEEE (2019)

4. Alexa: Alexa top websites. https://www.expireddomains.net/alexa-top-websites/ (2022), Accessed Sept 2022
5. AppsUK: How long does it take to get 1000 followers on instagram? https://apps.uk/how-long-1000-followers-on-instagram/ (2022) Accessed Jan 2023
6. Bailey, M., Dittrich, D., Kenneally, E., Maughan, D.: The Menlo report. IEEE Security & Privacy (2012)
7. Bedi, P., Sharma, C.: Community detection in social networks. Wiley Interdisc. Rev.: Data Mining Knowl. Disc. **6**(3), 115–135 (2016)
8. Boyd, D., Crawford, K.: Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inform. Commun. Society **15**(5), 662–679 (2012)
9. Brooker, P., Barnett, J., Cribbin, T., Sharma, S.: Have we even solved the first big data challenge?' practical issues concerning data collection and visual representation for social media analytics. In: Snee, H., Hine, C., Morey, Y., Roberts, S., Watson, H. (eds.) Digital Methods for Social Science, pp. 34–50. Palgrave Macmillan UK, London (2016). https://doi.org/10.1057/9781137453662_3
10. Campbell, C., Ferraro, C., Sands, S.: Segmenting consumer reactions to social network marketing. Europ. J. Market **38** (2014)
11. Conti, M., Gathani, J., Tricomi, P.P.: Virtual influencers in online social media. IEEE Commun. Mag. **60**, 86–91 (2022)
12. Conti, M., Pajola, L., Tricomi, P.P.: Captcha attack: Turning captchas against humanity. arXiv preprint arXiv:2201.04014 (2022)
13. Daugherty, A.: https://aigrow.me/follow-unfollow-instagram/ (2022) Accessed Oct 2022
14. Dayma, B., et al.: Dall-e mini (7 2021). https://doi.org/10.5281/zenodo.5146400,https://github.com/borisdayma/dalle-mini
15. De Cristofaro, E., Friedman, A., Jourjon, G., Kaafar, M.A., Shafiq, M.Z.: Paying for likes? understanding facebook like fraud using honeypots. In: Proceedings of the 2014 Conference on Internet Measurement Conference, pp. 129–136 (2014)
16. Del Vicario, M., et al.: The spreading of misinformation online. Proc. Natl. Acad. Sci. **113**(3), 554–559 (2016)
17. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 Ieee (2009)
18. for Developers, M.: Instagram api. hhttps://developers.facebook.com/docs/instagram-api/guides/insights (2021) Accessed Oct 2022
19. Dey, N., Borah, S., Babo, R., Ashour, A.S.: Social network analytics: computational research methods and techniques. Academic Press (2018)
20. Dittrich, D.: The ethics of social honeypots. Res. Ethics **11**(4), 192–210 (2015)
21. Face, H.: Keytotext. https://huggingface.co/gagan3012/k2t (2022). Accessed Oct 2022
22. Ferreira, N.M.: 300+ best instagram captions and selfie quotes for your photos. https://www.oberlo.com/blog/instagram-captions (2022) Accessed Sep 2022
23. Fisher, D., McAdam, A.: Social traits, social networks and evolutionary biology. J. Evol. Biol. **30**(12), 2088–2103 (2017)
24. Franke, R.H., Kaul, J.D.: The hawthorne experiments: First statistical interpretation. American sociological review, pp. 623–643 (1978)
25. Hagen, L., Keller, T., Neely, S., DePaula, N., Robert-Cooperman, C.: Crisis communications in the age of social media: a network analysis of zika-related tweets. Soc. Sci. Comput. Rev. **36**(5), 523–541 (2018)

26. Haqimi, N.A., Rokhman, N., Priyanta, S.: Detection of spam comments on instagram using complementary naïve bayes. IJCCS (Indonesian J. Comput. Cybern. Syst.) **13**(3), 263–272 (2019)
27. HQ, H.: How to get followers on instagram. https://www.hopperhq.com/blog/how-to-get-followers-instagram-2021/ (2022) Accessed Jan 2023
28. Hu, X., Tang, J., Liu, H.: Online social spammer detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 28 (2014)
29. Hub, M.: The state of influencer marketing 2021: Benchmark report. https://influencermarketinghub.com/influencer-marketing-benchmark-report-2021 (2021) Accessed Oct 2022
30. Infographic: Data never sleeps 5.0. https://www.domo.com/learn/infographic/data-never-sleeps-5 (2022) Accessed Oct 2022
31. Instagram: Reducing inauthentic activity on instagram. https://about.instagram.com/blog/announcements/reducing-inauthentic-activity-on-instagram (2018) Accessed Feb 2023
32. Instagram: Introducing new authenticity measures on instagram. https://about.instagram.com/blog/announcements/introducing-new-authenticity-measures-on-instagram/ (2020) Accessed Feb 2023
33. Jain, A.K., Sahoo, S.R., Kaubiyal, J.: Online social networks security and privacy: comprehensive review and analysis. Complex Intell. Syst. **7**(5), 2157–2177 (2021). https://doi.org/10.1007/s40747-021-00409-7
34. John, J.P., Yu, F., Xie, Y., Krishnamurthy, A., Abadi, M.: Heat-seeking honeypots: design and experience. In: Proceedings of the 20th International Conference on World Wide Web, pp. 207–216 (2011)
35. Karl: The 15 biggest social media sites and apps. https://www.dreamgrow.com/top-15-most-popular-social-networking-sites/ (2022) Accessed Sept 2022
36. Kim, R.E., Kotzé, L.J.: Planetary boundaries at the intersection of earth system law, science and governance: A state-of-the-art review. Rev. Europ., Compar. Int. Environ. Law **30**(1), 3–15 (2021)
37. Kreibich, C., Crowcroft, J.: Honeycomb: creating intrusion detection signatures using honeypots. ACM SIGCOMM Comput. Commun. Rev. **34**(1), 51–56 (2004)
38. Kuhn, S.: How to stop instagram spam? https://www.itgeared.com/how-to-stop-instagram-spam/ (2022) Accessed Jan 2023
39. Laurence, C.: Call to action instagram: 13 creative ctas to test on your account. https://www.plannthat.com/call-to-action-instagram// (2022) Accessed Sept 2022
40. Lavanya: How to avoid-stop spam comments on instagram posts? https://versionweekly.com/news/instagram/how-to-avoid-stop-spam-comments-on-instagram-posts-easy-method/ (2021) Accessed Oct 2022
41. Lee, K., Caverlee, J., Webb, S.: Uncovering social spammers: social honeypots+ machine learning. In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 435–442 (2010)
42. Lee, K., Eoff, B., Caverlee, J.: Seven months with the devils: A long-term study of content polluters on twitter. In: Proceedings of the international AAAI conference on web and social media. vol. 5, pp. 185–192 (2011)
43. Liu, J., Cao, Y., Lin, C.Y., Huang, Y., Zhou, M.: Low-quality product review detection in opinion summarization. In: Proceedings of the 2007 Joint Conference on Emethods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 334–342 (2007)
44. Macready, H.: The only instagram metrics you really need to track in 2023. https://blog.hootsuite.com/instagram-metrics (2022) Accessed Jan 2023

45. McClurg, S.D.: Social networks and political participation: the role of social interaction in explaining political participation. Polit. Res. Q. **56**(4), 449–464 (2003)
46. McCormick, K.: 23 smart ways to get more instagram followers in 2022. https://www.wordstream.com/blog/ws/get-more-instagram-followers (2022), accessed: Sep. 2022
47. Me, I.: How to get your first 1000 followers on instagram. https://www.epidemicsound.com/blog/how-to-get-your-first-1000-followers-on-instagram/ (2022) Accessed Jan 2023
48. Meyer, L.: How often to post on social media: 2022 success guide. https://louisem.com/144557/often-post-social-media (2022) Accessed Oct 2022
49. Moshchuk, A., Bragin, T., Gribble, S.D., Levy, H.M.: A crawler-based study of spyware in the web. In: NDSS. vol. 1, p. 2 (2006)
50. Murugan, N.S., Devi, G.U.: Detecting spams in social networks using ml algorithms-a review. Int. J. Environ. Waste Manage. **21**(1), 22–36 (2018)
51. Mushtaq, R.: Augmented Dickey Fuller Test. Mathematical Methods & Programming eJournal, Econometrics (2011)
52. OpenAI: https://openai.com/blog/chatgpt (2022) Accessed Mar 2023
53. Pereira, N.: 5 different tiers of influencers and when to use each. https://zerogravitymarketing.com/the-different-tiers-of-influencers-and-when-to-use-each/ (2022) Accessed Oct 2022
54. Petriska, J.: https://gist.github.com/JakubPetriska/060958fd744ca34f099e947cd080b540 (2022) Accessed Oct 2022
55. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020)
56. Rani, P., Shokeen, J.: A survey of tools for social network analysis. Int. J. Web Eng. Technol. **16**(3), 189–216 (2021)
57. Raponi, S., Khalifa, Z., Oligeri, G., Di Pietro, R.: Fake news propagation: A review of epidemic models, datasets, and insights. ACM Trans. Web **16**(3) (2022)
58. Richter, F.: Social networking is the no. 1 online activity in the u.s. https://www.statista.com/chart/1238/digital-media-use-in-the-us/ (2022) Accessed Sept 2022
59. Robertson, M.: Instagram Marketing: How to Grow Your Instagram Page And Gain Millions of Followers Quickly With Step-by-Step Social Media Marketing Strategies. CreateSpace Independent Publishing Platform (2018)
60. Sheikhi, S.: An efficient method for detection of fake accounts on the instagram platform. Rev. d'Intelligence Artif. **34**(4), 429–436 (2020)
61. Singh, A., Halgamuge, M.N., Moses, B.: An analysis of demographic and behavior trends using social media: Facebook, twitter, and instagram. Social Network Analytics, p. 87 (2019)
62. Smith, E.B., Brands, R.A., Brashears, M.E., Kleinbaum, A.M.: Social networks and cognition. Ann. Rev. Sociol. **46**(1), 159–174 (2020)
63. SocialBuddy: How often to post on social media: 2022 success guide. https://socialbuddy.com/how-often-should-you-post-on-instagram/ (2022) Accessed Oct 2022
64. Stallings, W., Brown, L., Bauer, M.D., Howard, M.: Computer security: principles and practice, vol. 2. Pearson Upper Saddle River (2012)
65. Statusbrew: Instagram algorithm 2022: How to conquer it. https://statusbrew.com/insights/instagram-algorithm/ (2021) Accessed Oct 2022
66. Stringhini, G., Kruegel, C., Vigna, G.: Detecting spammers on social networks. In: Proceedings of the 26th Annual Computer Security Applications Conference, pp. 1–9 (2010)

67. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
68. Tricomi, P.P., Chilese, M., Conti, M., Sadeghi, A.R.: Follow us and become famous! insights and guidelines from instagram engagement mechanisms. In: Proceedings of the 15th ACM Web Science Conference 2023, vol. 11, pp. 346–356. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3578503.3583623
69. Tricomi, P.P., Tarahomi, S., Cattai, C., Martini, F., Conti, M.: Are we all in a truman show? spotting instagram crowdturfing through self-training. arXiv preprint arXiv:2206.12904 (2022)
70. Vishwamitra, N., Li, Y., Hu, H., Caine, K., Cheng, L., Zhao, Z., Ahn, G.J.: Towards automated content-based photo privacy control in user-centered social networks. In: Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy. Association for Computing Machinery (2022)
71. Wang, G.,et al.: Serf and turf: crowdturfing for fun and profit. In: Proceedings of the 21st International Conference on World Wide Web, pp. 679–688 (2012)
72. Wang, Y.M., Beck, D., Jiang, X., Roussev, R.: Automated web patrol with strider honeymonkeys: Finding web sites that exploit browser vulnerabilities. In: IN NDSS. Citeseer (2006)
73. Webb, S., Caverlee, J., Pu, C.: Social honeypots: Making friends with a spammer near you. In: CEAS, pp. 1–10. San Francisco, CA (2008)
74. Xiao, Y., Jia, Y., Cheng, X., Wang, S., Mao, J., Liang, Z.: I know your social network accounts: A novel attack architecture for device-identity association. IEEE Transactions on Dependable and Secure Computing, pp. 1–1 (2022). https://doi.org/10.1109/TDSC.2022.3147785
75. Yang, C., Zhang, J., Gu, G.: A taste of tweets: Reverse engineering twitter spammers. In: Proceedings of the 30th Annual Computer Security Applications Conference, pp. 86–95 (2014)
76. Yegneswaran, V., Giffin, J.T., Barford, P., Jha, S.: An architecture for generating semantic aware signatures. In: USENIX Security Symposium, pp. 97–112 (2005)
77. Zhang, J., Zhao, Y., Saleh, M., Liu, P.J.: Pegasus: Pre-training with extracted gap-sentences for abstractive summarization (2019)
78. Zhang, W., Sun, H.M.: Instagram spam detection. In: 2017 IEEE 22nd Pacific Rim International Symposium on Dependable Computing (PRDC), pp. 227–228. IEEE (2017)
79. Zhang, Y., Zhang, H., Yuan, X.: Toward efficient spammers gathering in twitter social networks. In: Proceedings of the Ninth ACM Conference on Data and Application Security and Privacy, pp. 157–159 (2019)
80. Zhu, Y., Wang, X., Zhong, E., Liu, N., Li, H., Yang, Q.: Discovering spammers in social networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 26, pp. 171–177 (2012)