Published: 21 October 2020

TEEM'20: Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality
October 21 - 23, 2020
Salamanca, Spain

RESEARCH-ARTICLE

Automated image extraction from Instagram for social research: A technical and ethical exploration

MIGUEL VARELA-RODRÍGUEZ, University of Valladolid, Valladolid, Valladolid, Spain

MIGUEL VICENTE-MARIÑO, University of Valladolid, Valladolid, Valladolid, Spain

https://doi.org/10.1145/3434780.3436650
ISBN: 9781450388504

Automated image extraction from Instagram for social research

A technical and ethical exploration

Miguel Varela-Rodríguez
Department of Sociology and Social Work, Faculty of Commerce/University of Valladolid, Valladolid, Spain
miguel.varela@uva.es

Miguel Vicente-Mariño
Department of Sociology and Social Work, Faculty of Social, Legal and Communication Sciences/University of Valladolid, Valladolid, Spain
miguel.vicente@uva.es

ABSTRACT

The use of social media data in social research has grown exponentially since the early 2010s, with many social researchers incorporating some degree of social media analysis. Following the 2018 Cambridge Analytica controversy, most social media providers locked access to their users' data, leaving researchers with limited options to study the information in it. Scholars have taken different stands, advocating data policies that allow for critical research, achieving partnerships with data providers, or—sometimes—violating the Terms of Use of the platforms to access the information. Despite the limitations, great progress has been made using text-based data, while image-based methodologies remain limited, partly because of the lockout. This paper proposes a methodology for automated image extraction from Instagram, using Instalooter. It presents the necessary setup and steps to obtain images based on a series of search parameters before discussing the ethical implications of its potential use.

CCS CONCEPTS

• **Human-centered computing** → Collaborative and social computing; Collaborative and social computing systems and tools; Social networking sites.

KEYWORDS

Instagram, visual communication, public understanding of science, social media research, social media and social sciences, sociology of communication

ACM Reference Format:
Miguel Varela-Rodríguez and Miguel Vicente-Mariño. 2020. Automated image extraction from Instagram for social research: A technical and ethical exploration. In *Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'20), October 21–23, 2020, Salamanca, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3434780.3436650

TEEM'20, October 21–23, 2020, Salamanca, Spain

ACM ISBN 978-1-4503-8850-4/20/10…$15.00
https://doi.org/10.1145/3434780.3436650

1 INTRODUCTION

The analysis of social media and other sources of big data is steadily becoming part of the social researcher's toolkit, to great debate. Schäfer and van Es [1] refer to the emergence of algorithms and big (social media) data as an "era-defining" moment, and argue that terms like "digital humanities" will soon be outdated and its tools and methods simply incorporated into the broader social research curricula. Similarly, Gerwin van Schie, Westra and Schäfer [2] concede that "knowledge of such tools and practices is increasingly a requirement in academic hiring".

Examples of the practical application of social media data to social research abound, particularly in the field of political communications and cultural studies.

Using data from Twitter, WhatsApp and Facebook, Kligler-Vilenchik et al. [3] give an account of how political (de)polarisation takes shape, and how social media algorithms can influence the perception of political content. We have witnessed this dynamic for several years now in electoral processes across the Globe, with social media becoming key tools in campaigning. Work on Twitter [4] has shown how supervised, automated analysis can reduce the workload and timespan in approaching large pools of public data, allowing researchers to continue focusing on everyday life, of which social media is now an integral part.

For more than a decade now, the work of Lev Manovich and the Cultural Analytics Lab has impacted our understanding of visual elements in culture, whether those are films or selfies, and challenges practices that are commonplace in social research such as categorization [5, 6]. Smaller scale yet equally revealing work can be found in Munk et al. [7], who look at Instagram as part of the daily lives and offer three different perspectives on how to use Instagram for social research.

In 2018, the controversy following Cambridge Analytica's use of social media data highlighted the close relationship between data analysis, social behaviour, and policymaking, and resulted in sweeping bans in the access to social media data. Since then, access through scrapers has been largely banned, but not impossible.

Social researchers continue finding ways to use social media and its data to advance thinking and practice in different fields. A 2020 paper on the impact of Instagram on HIV risk perception makes explicit use of Instalooter [8], the Instagram scraper that will be discussed in this paper, highlighting the potential of automated image extraction to enhance health communication campaigns. Recent work to develop a dataset of COVID-19-related content on Instagram also makes explicit mention to scraping techniques [9], enabled through a partnership with Facebook, the parent company of Instagram.

Yet the use of image-driven media, such as Instagram or Snapchat, remains an emerging field compared to the more general use of text-driven media. Efforts often either rely on manual identification and extraction of images, focus on the accompanying texts (instead of the images), or fall into ethical and legal pitfalls.

This paper discusses a methodology for the semi-automated obtention and storing of images from Instagram, based on the tool Instalooter, an image scraper for Instagram and used in previous work such as Nobles et al. [8].

We first provide a list of the tools necessary to approach the extraction of images from Instagram, and then move into the application of Instalooter, using two fictional examples to guide the reader. The methodology presented is, admittedly, rather artisanal, and it does not reach the automation of machine-learning and other modern techniques.

While this paper may inform specific applications for image extraction from Instagram, it should be read as a critical review of the use of the tool, both from a technical and an ethical point of view.

2 STEP 1. TOOLS REQUIRED AND SETUP

The purpose of this step is to set up all the necessary tools prior to implementation. The tools proposed are publicly available and free of charge, except for Microsoft Excel, which is available to many social researchers through their departments. They require some knowledge of programming and being comfortable with managing large datasets but are generally user-friendly and well-documented.

2.1 Python (Anaconda) and Pip

Python is a programming language many social researchers are familiar with as an alternative to R for statistical analysis. It can be downloaded and installed using Anaconda, an easy-to-access toolkit that contains several data science tools useful for social researchers. While there are alternative ways to install Python, the Anaconda package is perhaps the most user-friendly and is available from its official webpage: https://www.anaconda.com/products/individual

Installing Anaconda will ensure Python is correctly setup.

Meanwhile, Pip is a "package installer" for the Python language. It allows users to download and install additional tools to be used through Python. Pip can be used to download and install Instalooter, a web scraper for Instagram.

2.2 Instalooter

As per its online description, "Instalooter is a program that can download pictures and videos from any profile or hashtag on Instagram" [10]. It thus fits the description of a "web scraper", better understood as "bits of software code that makes it possible to automatically download data from the web" [11]. At the time of writing, Instalooter is available (despite breaching the Terms of Use of Instagram) and can be downloaded using Pip:

- Launch Command Prompt.
- Launch "python".
- Enter this code: pip install.package instalooter.

Further documentation can be found on the homepage of the developers: https://instalooter.readthedocs.io/en/stable/usage.html#usage

3 STEP 2. DEFINING THE SEARCH QUERY THROUGH OBSERVATION

The purpose of this step is to prepare a search query that can be implemented using Instalooter while observing ethical guidelines. It can be equated to a design and safe data exploration phase [2].

As its name suggests, Instalooter facilitates a process of "looting" on Instagram: with basic knowledge of the Python language, users can program an automated, mass download of publicly available images from Instagram that meet a series of parameters previously defined. It does not require an existing Instagram account or application, greatly facilitating the obtention of data for social research.

Marres & Weltevrede describe the process of scraping as "a series of steps in which formatted data is derived from an informational mess" [11]. That is: scraping can help categorize, order and funnel data for deeper analysis in a world where data is everywhere. In their exposition, Marres & Weltevrede explore scraping as a tool that can help researchers "adopt analytic categories that have acquired saliency" [11]. On Instagram, those categories will often take the shape of hashtags.

While not failproof, hashtags are a useful way to narrow down a search on Instagram, especially when they have been previously identified by the researchers. Whereas manual extraction (through screenshots, for example) is likely to be impacted by the salience of images in the researchers' feeds, automated extraction allows for the bulk of images that carry the given hashtag to be downloaded. In parallel, researchers may include additional accounts or hashtags, and define a timeline to further narrow down the search.

Although it may seem intuitive to first conduct an automated exploration and then extract the categories/hashtags, it is worth reviewing the implications of such an approach.

Firstly, the sheer volume of images on Instagram makes it unadvisable to attempt mass-downloads without defining clearly delimited queries, as it may lead to analysis paralysis. Secondly, there would be little justification in conducting a mass download outside the scope of the work.

Much like in "traditional" social research, this early phase necessitates careful observation and thorough design to avoid putting users at risk. That is, researchers ought to inhabit the online space, follow the accounts that are immediately relevant to their subject, detect hashtags used in their context of study, and only then consider downloading images. This follows the design proposed by Gerwin van Schie, Westra and Schäfer [2]: to research social media, you must first design the process and conduct a safe data exploration, anticipating potential ethical conundrums.

To illustrate this step, two research scenarios are defined:

- *In political and peace and conflict studies*: following the arrest of an opposition party leader on 15 June 2020, the opposition party calls for widespread protests using their Instagram account. Through it, the party posts daily images related to the party leader and the protests underway. Researchers wish to study the images produced before and after by the party, so they choose to define a timeline of two weeks before and two weeks after the arrest.
- *In public health and education*: as the academic year is set to begin at the University of Valladolid following the COVID-19 lockdown, researchers are interested in observing the images

uploaded by students of their lecture halls and facilities, and how they visualise their faculty. They use a series of hashtags previously identified to be used by students to demand action in the wake of restrictive measures, mainly "#LaUVaNoResponde" ("the University of Valladolid doesn't respond").

While supporting observation with automated downloads can help speed up the process and reduce the transaction cost of manually reviewing the images and capturing screenshots, it does not do away with the key challenges in the bias present when selecting the keywords to be studied [12]. As a potential link with qualitative approaches, digital ethnography (netnography) could turn into a valuable practice as a guide to polish and refine our search queries, identifying those traces that are more useful to accomplish our research purposes.

Once the parameters have been defined, the search query seems clear:

- Scenario 1, interested in a specific account: all images by the Instagram account associated to the movement between 1 June 2020 and 30 June 2020.
- Scenario 2, interested in a hashtag: all images containing the hashtag "#LaUVaNoResponde" between 1 September 2020 and 30 September 2020.

4 STEP 3. OBTAINING THE IMAGES AND RELEVANT METADATA

4.1 Working with the images

With the search query ready and the code adapted, building the Instalooter code is straight-forward. To launch Instalooter:

- Open Anaconda Prompt
- Type: CD [full path of the folder where you wish to download the images]

Researchers can define a specific target folder to store the images in their computer, while additional metadata can be attached to the filename (such as account name or date). Metadata can also be added to the filename, which makes identification and managing of images easier.

Before attaching any account-related information, however, it is worth exploring all issues surrounding data protection and privacy, as well as existing guidelines on ethical research. In most cases, the username, date, and ID will only be relevant for verification purposes (i.e., in case other researchers wish to replicate the research process) while they could pose a risk to users.

Following the two scenarios above, only one line of code would be necessary to perform the search for images. These would be the lines executed once Python and Instalooter are active:

```
instalooter user USERNAME -t 2020-06-01:2020-06-30 -T{date}{id}
instalooter hashtag LaUVaNoResponde -t 2020-09-01:2020-09-30 -T{date}{id}
```

Executing the code from Python will begin a download process for all images fitting the search, which may take anywhere from two minutes to over an hour, depending on the number of images available. This is the process known as "scraping".

Images downloaded using Instalooter are automatically stored with the unique ID of the image on Instagram attached to the filename. This facilitates identification but poses a series of ethical challenges as the image could be traced back to its user. Such information can be removed by defining the format of filenames through Instalooter (documentation available online) or removing them post-download.

Images are downloaded in JPEG format, and researchers may need to manually review that they are indeed relevant to the study. Depending on the aim, topic and sensitivities surrounding the project, images containing personal information should be removed to avoid exposure of vulnerable subjects.

4.2 Working with metadata only

Alternatively, researchers may decide not to download images directly, but to simply dump the associated metadata in a text file. This is useful when the intention of the research is to analyze the volume of images surrounding a hashtag or event, as opposed to their actual visual content. There is some consensus that it is the volume of images that captures the patterns of social communication most relevant to social researchers [3].

In the cases explored above, the code to extract the metadata without the images would look exactly as in the previous example, with an added "-D" at the end:

```
instalooter user [USER] -t 2020-06-01:2020-06-30 -T{date}{id} -D
instalooter hashtag LaUVaNoResponde -t 2020-09-01:2020-09-30 -T{date}{id} -D
```

This will dump the metadata for all images found with the search query defined by the researcher into a text file.

While this approach holds value, there are cases where the individual image is important, particularly when social researchers aim to understand meanings and discourses.

5 TECHNICAL CONSTRAINTS

While the method described above may facilitate data collection and sorting, it is not without technical limitations. It provides a shortcut to obtaining images, but it does not help the methodological limitations of defining the "right" hashtags to monitor. These are questions explored by Hand [12] or Schäfer & van Es [1] and cannot be addressed without a review of the ethical design of the research.

Meanwhile, recent developments in access to Facebook and Instagram data suggest the code for Instalooter is likely to stop working sooner than later. In the wake of the Cambridge Analytica-motivated policy changes, access to data is reduced and often unpredictable [13].

In addition, the Instalooter code does not allow for complex queries, such as searches for images that contain combinations or intersections of hashtags. These searches would be useful to filter irrelevant content. Similarly, at the time of writing it is not possible to conduct a search by location, which reduces its usefulness in location-specific research.

Lastly, while limited, the method does require some degree of coding knowledge and familiarity with digital tools for research, which are not necessarily part of all social research teams.

Similar tools, such as Instaloader or Instagram PHP Scraper, face similar challenges. Alternative methods may be explored, especially achieving a partnership with Facebook for safer access to images. Commercial options exist, too, such as Meta Eyes, which incorporates image recognition, or Vurku, but their reliability is yet to be tested in academic research.

6 ETHICAL CONSTRAINTS

If there is a key question mark lurking over this method, it is the ethical challenges that it poses. While scrapers have been around for over a decade, social researchers are often unfamiliar with them and perceive them as unsafe practice, "mostly closely associated with illegal practices" [11]. Indeed, article 10 of the Instagram Terms of Use specifically bans the use of such techniques [14] and, while the 2018 update to the terms does not make explicit mention to scraping, it is safe to assume this technique will continue to be banned.

Beyond the legality of using scrapers for research purposes, it is also worth considering questions around consent and ownership, important to online ethnography and which will become all the more pressing when dealing with sensitive issues like health, living conditions, education and other topics where users are likely to share personal images.

The ethics of studying social media data have been a subject of debate in the last decade. The discussion is often compounded not only of what is right (or safe) and what is not right (or not safe) to download and study, but on the epistemological discussion of where social media fits within social research.

This discussion is present in Marres and Weltevrede, who concede that incorporating social media data analysis in the curricula of social research may open the space to "entities that are in some respects alien to the context of academic social research" [11]. Markham [15] points to the challenges that such opening could entail, discussing how basic considerations that scholars are familiar with (such as vulnerability, consent or ownership) may not be part of the lexicon of data analysts approaching social research.

Throughout, researchers are advised to exercise caution and follow ethics guidelines. Markham and Buchanan's [16] framework is firm ground on which to stand. It provides the conceptual basis as well as clear direction on when approaching social media data can put research participants at risk and encourages researchers to incorporate decision-making as a deliberate part of their work. That is: assess the potential for identificacion, evaluate the vulnerability of the subject, consider rights and benefits, judge the sensitivity of the topic, and only then consider using scrapers like Instalooter.

In addition to Markham's and Buchanan's approach, Gerwin van Schie's, Westra's and Schäfer's three-step process to social media research provides robust support [2]. Preceding each step of the research process with what they call a "moment of ethical reasoning" can help scholars working with social media data avoid legal pitfalls and—more importantly—avoid harm to research subjects or participants.

Which leads to the last consideration, the very word "participant". Both in the conception of the researcher and of the subjects as participants, Markham [15] advocates for a review of the concept of participation in social media. When dealing with thousands of social media accounts, it seems unlikely that subjects can be considered participants.

7 CONCLUSIONS

In their reactions to the Cambridge Analytica-motivated data lockdown, Bruns [17] and Rogers [13] explore different options for social researchers to continue building an understanding on how social media affects the lives of millions of people. Faced with new constraints, Bruns highlights three different options for scholars: walking away, advocating for change. . . or breaking the rules.

This paper provided an overview of one of the existing methods available to researchers. We walked the reader through the steps to access Instagram images, the tools required, and the implications of approaching this work.

While publicly available data is increasingly kept away from critical research, social researchers are increasingly faced with the need (both instrumental and in the form of market pressure) to become acquainted with the tools to understand it.

We consider the use of Instalooter, a publicly available tool that uses the Python programming language, to set up a research process. We begin with a design phase to establish the search query and then move to the process of obtaining said images.

This paper also contributed to the broader discussion on online research and ethics, concluding that the use of Instalooter, beyond being a breach of the Terms of Use of Instagram, poses serious challenges in the conception of social research and its relationship with its participants. We proposed that researchers approach this method only after they have conducted a thorough review of their design process, following existing guidelines in ethical design in online research.

REFERENCES

[1] Schäfer, M.T. and Es, van, K. eds. 2017. The Datafied Society. Studying Culture through Data. Amsterdam University Press.

[2] Gerwin van Schie, I. *et al.* 2017. Get Your Hands Dirty: Emerging Data Practices as Challenge for Research Integrity. The Datafied Society. Studying Culture through Data. M.T. Schäfer and K. Es, van, eds. Amsterdam University Press.

[3] Kligler-Vilenchik, N. *et al.* 2020. Interpretative polarization across platforms: How political disagreement develops over time on Facebook, Twitter, and WhatsApp. Social Media Society. (2020).

[4] Arcila-Calderón, C. *et al.* 2019. Análisis distribuido y supervisado de sentimientos en Twitter: Integrando aprendizaje automático y analítica en tiempo real para retos de dimensión big data en investigación de comunicación y audiencias. Empiria. Revista de metodología de ciencias sociales. 0, 42 (Jan. 2019), 113–136. DOI:https://doi.org/10.5944/empiria.42.2019.23254.

[5] Manovich, L. 2018. Can We Think Without Categories? Digital Culture & Society (DCS). 4, 1 (2018), 17–28.

[6] Manovich, L. 2016. Instagram and Contemporary Image.

[7] Munk, A.K. *et al.* 2016. (Re-)Appropriating Instagram for Social Research: Three Methods for Studying Obesogenic Environments. Proceedings of the 7th 2016 International Conference on Social Media & Society (New York, NY, USA, Jul. 2016), 1–10.

[8] Nobles, A.L. *et al.* 2020. Automated image analysis of instagram posts: Implications for risk perception and communication in public health using a case study of #HIV. PLOS ONE. 15, 5 (May 2020), e0231155. DOI:https://doi.org/10.1371/journal.pone.0231155.

[9] Zarei, K. *et al.* 2020. A First Instagram Dataset on COVID-19. arXiv:2004.12226 [cs]. (Apr. 2020).

[10] InstaLooter — InstaLooter 2.3.4 documentation: 2019. https://instalooter.readthedocs.io/en/stable/index.html. Accessed: 2020-09-14.

[11] Marres, N. and Weltevrede, E. 2013. SCRAPING THE SOCIAL?: Issues in live social research. Journal of Cultural Economy. 6, 3 (Aug. 2013), 313–335. DOI:https://doi.org/10.1080/17530350.2013.772070.

[12] Hand, M. 2016. Visuality in Social Media: Researching Images, Circulations and Practices. The SAGE Handbook of Social Media Research Methods. SAGE

Publications Ltd. 215–231.
[13] Rogers, R. 2018. Social Media Research After the Fake News Debacle. University of Salento.
[14] Terms of Use •Instagram: 2013. https://www.instagram.com/about/legal/terms/before-january-19-2013/. Accessed: 2020-09-14.
[15] Markham, A.N. 2013. Fieldwork in Social Media: What Would Malinowski Do? Qualitative Communication Research. 2, 4 (Dec. 2013), 434–446. DOI:https://doi.org/10.1525/qcr.2013.2.4.434.
[16] Markham, A.N. and Buchanan, E. 2017. Research Ethics in Context: Decision-Making in Digital Research. The Datafied Society. Studying Culture through Data. M.T. Schäfer and K. Es, van, eds. Amsterdam University Press.
[17] Bruns, A. 2019. After the 'APIcalypse': social media platforms and their fight against critical scholarly research. Information, Communication & Society. 22, 11 (Sep. 2019), 1544–1566. DOI:https://doi.org/10.1080/1369118X.2019.1637447.