Studies in Communication Sciences (2024), pp. 1–16

Strategies and challenges for constructing and collecting visual corpora from image-based social media platforms

Yuliya Samofalova, Université Catholique de Louvain, Institut Langage et Communication, Belgium
yuliya.samofalova@uclouvain.be

Abstract

Visual elements play an important role within the multimodal nature of social media (Pearce et al., 2020). A growing body of research has focused on the analysis of still and moving images from different social media platforms from various perspectives of communication and media studies (Hautea, Parks, Takahashi, & Zeng, 2021; Li & Xie, 2020; Veum & Undrum, 2018). Although the aforementioned studies describe visual data collection, their principal focus does not rely on this collection, but on data analysis. Little attention has been paid to the challenges of collecting visual datasets (Highfield & Leaver, 2016). In this paper, I propose a methodological overview of several strategies for collecting large corpora of visual data from image-based social media platforms. Provided with exemplary publications, I review five strategies for collecting visual corpora: hashtag-based, account-based, metadata-based, random sampling, and mixed approach. Lastly, I present a case study with my own mixed approach to the collection of visual data from Instagram. Considering the usage, advantages and limitations of each strategy, the article will contribute to the developing science of social media research. I believe that a literature analysis of visual data collection strategies and a provided case study can help researchers optimize visual data collection from image-based social media.

Keywords

visual corpus, social media, data collection, Instagram

1 Introduction

Recent findings report that more than 50 billion photos have been shared on Instagram since its launch in 2010 (Aslam, 2023) and more than 500 hours of video were uploaded to YouTube every minute in 2022 (Statista, 2022) These facts can be referred to as the idea of social media "visual turn" (Gibbs, Meese, Arnold, Nansen, & Carter, 2015), which has influenced academic research in the field of communication and media studies and introduced new methods, challenges, and opportunities for analyzing society by examining social media data.

Social media provides an "easy and convenient access to large quantities of data," including visual content such as images, videos, and gifs (Nau, Quan-Haase, & McCay-Peet, 2022, p. 21). The research on images from social media has been rapidly developing since the early 2000s (Fung et al., 2019). Indeed, many studies have focused on the analysis of images from Flickr, Twitter, and Instagram (Samani, Guntuku, Moghaddam, Preoţiuc-Pietro, & Ungar, 2018; Trillò et al., 2021). Concurrently, many other studies investigated the videos from YouTube (Allgaier, 2019; Waters & Jones, 2011). These works contributed to the development of visual methodologies and social media studies by addressing theoretical and methodological issues. However, they do not rely on collecting the corpora from social media. In addition to this, most of the corpus-based social media research relies on textual data collection (Al-khateeb & Agarwal, 2019) or data retrieved from blogs (Weller, Bassalo, & Pfaff, 2018). Hence, I argue more research should be done within the field of visual data collection from social media platforms.

Furthermore, I propose a theoretical and methodological reflection on the existent approaches to visual corpora collection from more recent image-based social media platforms, specifically Instagram. The objective of our contribution is twofold. First, I synthesize the works within communication and

https://doi.org/10.24434/j.scoms.2024.01.3881

media studies and their approaches to visual data collection. Second, I present a case study as an example of a mixed approach to collecting visual data from Instagram and some tools that can be efficiently used to retrieve this type of data. Finally, I delineate the limitations and future perspectives of the research.

2 Visual corpus from image-based social media

The visual and the media play a significant role in social, political, and economic life which cannot be ignored (Loizos, 2000). As data-driven disciplines, communication and media studies use visuals as prime data. In particular, a large volume of visual data is generated and provided within image-based social media (Pearce, Niederer, Özkula, & Sánchez Querubín, 2019). These platforms use visuals, still and moving images, as predominant types of content and are also referred to as visual social media (Leaver, Highfield, & Abidin, 2020). For instance, Flickr, Instagram, Pinterest, Snapchat, TikTok, Tumblr, and YouTube can be considered image-based social media, since it is not possible to produce a post within them without opting for a visual mode.

Following the works of Schroeder (2018) and Lindgren (2018), I define social media data as material generated by users of social media platforms, which exists primarily before analysis, and can be collected and devised into different units of analysis. In communication research, this material is not only used as observation data but also for corpus creation. A corpus, according to Barthes (1967, p. 96), is "a finite collection of materials, which is determined in advance by the analyst, with (inevitable) arbitrariness, and on which he is going to work." In the research field on social media, the materials for the analysis can be user profiles, posts, or the visual or verbal components of a post. By visual components, hereafter referred to as visuals, I understand visual social media data (O'Halloran, Chua, & Podlasov, 2014), i. e., the still and moving images, which are generally posted in a separate block of post's content and whose roles are significant within the messages of a social media platform. Visual components can be photographs, graphic illustrations, posters, screenshots, gifs, as well as short and long videos. Although visual and verbal components (post's caption) are generally displayed in separate blocks on the website and in the application, the visual component can also include various text on screen elements, such as subtitles to the video, title, text, or infographics. In this context visual component of a social media post can be named multimodal. In this article, I use the notion of visual component in general terms highlighting the particular block or part of the post's visual content.

Following the literature analysis, I distinguish two main typologies of visual corpora. The first typology relies on the type of visuals included in the corpus, namely image, video, or multimodal document. While image or video corpora include only one type of visual, multimodal corpora involve any type of visual alongside textual components from social media posts. On the other hand, the second typology of visual corpora is based on the number of platforms involved in the data collection process which can be single-platform or multi-platform visual corpora. Hence, single-platform studies involve the collection and analysis of visual corpora from one social media platform, while multi-platform research involves the collection of datasets from several social media and is a developing field of media and communication sciences. Single-platform studies, on the contrary, are the established norm for social media analysis (Rogers, 2017). However, the research on single-platform visual data is mostly focused on their analyses, not on the collection process. Hence, in this paper, I propose some strategies for a single-platform multimodal corpora collection, i. e., visuals and texts posted on one social media platform. Visuals here are collected as a part of multimodal research projects and represent primary data for further analyses.

3 Methodological strategies of data collection

In this part, I propose a literature analysis of various approaches to visual data collection

from social media which may be considered the most frequently used by researchers in communication since 2015.

Several studies reviewed social media data collection with particular focus on Facebook (Abdesslem, Parris, & Henderson, 2012), or on the challenges of data collection for social media analytics (Stieglitz, Mirbabaie, Ross, & Neuberger, 2018). In communication research, Mayr and Weller (2016) suggested four approaches to collecting social media data from Twitter and Facebook: based on user accounts, topics and keywords, metadata, and random samples. Although these methodologies were mainly developed for the analyses of textual content, they are also relevant for visual data collection. I follow up on the work of Mayr and Weller and adapt their classification of social media data collection strategies to visual data. I present five approaches to data collection: hashtag-based, account-based, metadata-based, random sampling, and mixed approach. While the first four approaches are adapted from the work of Mayr and Weller, the selection of a mixed approach resulted from the observations and synthesis of recent academic works in the communication field. Several studies have emphasized in particular the importance of combining strategies for data collection to construct a dataset that would correspond to the objectives of the project (Ferchaud, Grzeslo, Orme, & LaGroue, 2018; Hartika, Pawito, & Utari, 2022; Veum & Undrum, 2018). Therefore, in this article, I see a high potential in presenting a mixed approach as a separate method to visual corpora collection from social media, which has its own specifics.

Our synthesis of the advantages and drawbacks of each strategy, outlined in Table 1, provides examples of research aims and methods in communication and media studies that could be applied to this type of data. The methods from Table 1 are related to investigating either visual or multimodal social media data from the perspective of communication disciplines. Although this is an up-to-date table it has to be completed over time. In addition, the presented strategies do not involve users' active participation in data collection; data is either manually or automatically gathered from social media. In the following sections, I detail the five strategies.

3.1 Hashtag-based

A hashtag-based strategy is one of the most frequent approaches to visual data collection from image-based social media. Hashtags are used as keywords to attract the attention of particular audiences to a given subject. This contributes to their role of "content-organizing elements" (Rogers, 2017, p. 95). According to Rogers (2017), hashtags can be used efficiently for a remote event analysis of elections, revolutions, social causes, disasters, subcultures, citizen movements, and the like.

Hashtag-based data collection is an elaborated method to collect visual corpora from social media: Many studies in communication can be used as examples for the investigation of hashtag-based data. They focus on content analysis of images from Instagram (Y. Kim, Song, & Lee, 2020) and TikTok videos (Zeng & Abidin, 2021), thematic framing analysis of Instagram posts (E. Lee & Weder, 2021), critical discourse analysis of memes from Twitter and Instagram (Boling, 2020), sentiment and theme analysis of the videos from TikTok (Rutherford et al., 2022), sentiment analysis of the images from Flickr and Twitter (Hassan et al., 2022), and visual cross-platform analysis to investigate platform vernaculars of multiple social media platforms (Pearce et al., 2020). Each social media is built with different algorithms: the use of hashtags may differ according to the platform as shown by Bossetta (2018). Thus, these differences should be considered by researchers before the data collection and analyses.

Since all selected posts are thematically focused, this strategy allows the observation of a target group's communication. It can be used to investigate various types of activism on social media: structures of social movements, e. g., #FridaysForFuture; messages and visuals related to particular events proposing their hashtags, such as #COP26; social issues which can also be emphasized by hashtags on social media, as #nuclearenergy. Depending on the affordances of the collection tools and the time scale of a project, hashtag-based data can be a suitable option for longitudinal research: as in Ahrens' et al.

Table 1: Strategies of data collection

Strategy	Can be applied to study	Research methods to analyze visual or multimodal corpora with exemplary publications	Benefits highlighted within the example studies	Limitations identified within example studies
Hashtag-based	hashtag activism, structure of social movements, particular events, social problems	content analysis, thematic framing analysis, critical discourse analysis, sentiment analysis, visual cross-platform analysis	many works involve the strategy; the posts are thematically focused; provides an overview of target group communication; can be used for longitudinal research	some posts can include duplicates; data requires extra sorting: by language, eliminating posts by bots or posts not related to the hashtag theme; certain connections cannot be explained by hashtag data only; cannot characterize the target group or socio-demographic situation
Account-based	self-representation, virtual identity, social identity, platform affordances	content analysis, framing analysis, multimodal critical discourse analysis	many works follow the strategy; provides an overview of individual or organizational communication; can be used for longitudinal research; lists of particular accounts can be pre-generated by some tools	may require the use of additional research methods, e. g., online ethnography, interviews
Metadata-based	platform vernaculars and affordances, visual content within a particular geographical area, trends	automated content analysis, computational metrics and qualitative analysis	might be used for longitudinal research	not all social media permit to collect the posts based on their metadata only; unstable nature of posts' metadata
Random sampling	personality traits prediction, modalities of the content, social media usage patterns, platform vernaculars	content analysis, cross-modal and cross-platform analysis	the dataset can already exist and be easy to download, which reduces time of data collection	not every research subject can be a part of an already-created random sample; difficult to achieve sufficient level of data representativity
Mixed approach	a particular campaign within one account, trends within a platform, trends considering a particular theme within a location	content analysis, multimodal discourse analysis, framing analysis, multimodal critical discourse analysis	contextualized and more precise dataset	can be a long-term process of obtaining and sorting data; outcomes may depend on platform specifics; few tools available

(2022) longitudinal content analysis to investigate fitness-related images from Instagram.

Despite providing an overview of a particular theme or event, a hashtag-based strategy of visual data collection does not enable the characterization of a whole target group or a sociodemographic situation. This is mainly due to the abundant number of visuals shared on social media and the limitations of the data collection tools. Nevertheless, preliminary observations and careful choice of hashtags, e. g., using special tools for selecting hashtags, can help overcome this limitation. Another limitation of this method of data collection lies within the construction of a rigorous dataset, since the collected messages may be not relevant to the research theme, or include posts in different languages from different countries outside the target corpus. Therefore, the data should be well-selected and additionally sorted before the analyses, since it can include multilingual elements, duplicates,[1] bot-generated content, or messages not directly related to the theme of the research.[2] In this case, automated collection of the dataset should be complemented by a manual revision.

1 In the cases when users post the same message multiple times or repost someone's message.

2 When the top-rated hashtags are used for any post's message in order to gain more engagement.

3.2 Account-based

A shared methodology for the collection of visual data from social media is based on users' accounts. This strategy is mostly employed when the object of the research is an account or a group of accounts. Such work is usually either personality-focused or organization-focused. The data collected from a particular account can be used to investigate a user's virtual identity and self-representation, social identity, or platform affordances, in particular for multi-platform studies. Numerous studies used an account-based strategy to collect visual data. They specifically focused on content analysis of Instagram and Twitter posts (Gruzd, Lannigan, & Quigley, 2018), visual content analysis of Instagram data (Moffitt, 2024), content analysis of Twitter, Facebook, and YouTube posts (Smith, Fischer, & Yongjian, 2012), framing analysis (Molder, Chen, Clemmons, & Lakind, 2021), and multimodal critical discourse analysis of posts from Instagram (Mapes & Ross, 2022). Most of these works confirm that additional qualitative methods, including interviews or observations, can be used to complement the results from analyzing visuals on social media. This can be a possible limitation for certain research projects.

Account-based strategy is an elaborate method to collect data from social media. As hashtag-based, this strategy can be used for longitudinal research. It can also provide an overview of individual or organizational communication. Another advantage of this method is that in some cases, the lists of accounts can already be provided by research groups or platforms for collecting and analyzing this data, e. g., *CrowdTangle*. In other cases, it can be effortful to identify accounts of particular group members and, as Mayr and Weller (2016, p. 112) state, "decisions will have to be made about who to include or not." Indeed, the study of Cougnon, de Viron, and Watrin (2022) has illustrated the challenges of automated sampling of targeted Twitter accounts according to their location, which can also be relevant to account sampling on image-based social media.

3.3 Metadata-based

Some research projects do not specify requirements related to a particular account or theme: thus, since they address the platform vernaculars, any type of visual content from a particular location, or the most trending posts for a given period, becomes acceptable. In this case, posts' metadata can be used as a "suitable entrance point," as argued by Pearce et al. (2020, p. 168) following the work of Rogers.

Metadata is a very flexible component of a social media post since this information can be updated every minute. The date, time, location, numbers of likes, comments, or shares can be considered as posts' metadata when the account provides access to it. Therefore, the research results will always depend on the metadata gathered at the time of collection. However, much effort should be made with computational sciences to collect visual data based on its metadata due to the limitations of existing tools for data collection.

In comparison to the research based on data collected with the help of hashtags or accounts, the number of studies focusing on visual corpora built around its metadata is relatively small. However, some prominent works can be used as example studies for future investigations. For instance, focusing on the popularity of YouTube videos, Rieder, Matamoros-Fernández, and Coromina (2018) investigate how the content is influenced by platform vernaculars, whereas the work of Tifentale and Manovich (2015) on automated analysis and visualization relies on evaluating Instagram images from 13 cities.

In addition to these works, I suggest that visual data collected from a particular location may be used for longitudinal research. For instance, it can be applied to investigate the evolution of themes communicated within a particular region or country. It should be considered that location-based metadata may not be completely reliable and thus might not contribute to the generalization of results. This is due to the flexible nature of location-based metadata: while some users may not indicate it, others could intentionally indicate a wrong geographical position. Hence, this type of visual corpus should be carefully verified and filtered if necessary.

Based on engagement metrics, researchers can observe the trends within the platforms and analyze platform vernaculars

and affordances. However, this strategy is more frequently used in combination with hashtag- or account-based strategies and is described in detail in section 3.5.

3.4 Random sampling

Random sampling can be particularly appealing for data analytics since sampled data preserve a number of statistical properties of the global set (H. Kim, Jang, Kim, & Wan, 2018). Random samples can be efficient to predict personality traits and to observe the modalities of the content, patterns of social media use, and platform vernaculars. In comparison to the previous strategies, less research exists on a random sampling of visual data. On the contrary, most of the works in communication research apply this strategy to the analysis of textual data (Su et al., 2022; Veltri & Atanasova, 2017). Nevertheless, random sampling can also be applied for collecting and analyzing visuals. The work carried out by Samani et al. (2018) focuses on the data collected from 300 random accounts on Flickr. Random samples in visual communication research have been used to conduct content analysis (León, Negredo, & Erviti, 2022), cross-modal and cross-platform analysis (Samani et al., 2018), and multimodal critical discourse analysis (Veum & Undrum, 2018). However, the majority of these studies apply random sampling to prepare visual corpora for analysis. In other words, the visuals are randomly selected from existing datasets.[3]

Random sampling is a developing strategy for visual data collection. One of its assets lies within the existing open-source visual datasets from social media on various themes available for the researchers. However, not every research subject is a part of already collected corpora. Another limitation of this strategy lies within the affordances of collection tools from image-based social media, which for now do not allow a collection of randomly selected visuals or posts. Additionally, it can be difficult to construct a representative corpus in particular, if more recent data is needed.

3 This process is related to mixed approach and described in the next section.

3.5 Mixed approach

A mixed approach to visual data collection can include a combination of two or more of the approaches described above. On the one hand, this strategy is used to explain certain connections within the corpus, for example, the reasons for applying a particular hashtag within one account. On the other hand, this approach is frequently used for post-collection sampling or filtering. Various types of mixed approaches to visual data collection can be identified, according to the order of application: hashtag–account, hashtag–location, hashtag–likes, hashtag–random sampling, account–hashtag, account–metadata, account–random sampling, etc. In the following paragraphs, I discuss some prominent studies which employ the most frequent mixed approaches to visual corpora creation.

Studies by Hartika et al. (2022) and Pramana, Utari, and Naini (2021) serve as an illustration of the account–hashtag strategy. The researchers examine the images posted by a particular Instagram account containing a selected hashtag. This strategy contextualizes the visual corpus of a project and can provide an outlook of a particular campaign of the investigated organization.

Hashtag–likes is another widely used mixed strategy in communication research. In particular, the data is firstly collected from a hashtag, then sorted according to the number of likes. Such strategy is used to identify trends and their origin, to examine technical features of the platforms (Hautea et al., 2021), or to identify themes (Shi et al., 2022) and particular patterns within the posts (Y. Lee, Huang, Blom, Schriner, & Ciccarelli, 2021).

Account–random sampling is an effective strategy to select random videos or images from a particular account for content analysis. Specifically, this strategy is worthwhile when the account presents a lot of content. In the work of Ferchaud et al. (2018, p. 91), "one video per month for a period of 24 months […] was selected using a random number generator." This strategy allows the comparison within selected accounts and provides a certain level of data representativity as shown in the study by Li and Xie (2020).

The article by Veum and Undrum (2018) illustrates the hashtag–random sampling approach. The authors randomly selected 100

selfie images containing the hashtag *#selfie* from Instagram. In order to provide relevant data for further multimodal analysis, the posts were additionally filtered to correspond to some predefined criteria.

To summarize, various mixed approaches to collect visual data from image-based social media can be identified. They are used to investigate particular campaigns of a selected account, trending content on a platform, and trends within a specific theme of a selected geographical area. The research methods that can be applied to this data are content analysis, multimodal discourse analysis, framing analysis, and multimodal critical discourse analysis. A combination of various strategies allows researchers to build a contextualized dataset, therefore, optimizing the data analysis process. Despite these advantages, the outcomes of a mixed approach to social media data collection can largely depend on platform specifics, especially if demographic characteristics are considered. In addition to the number of posts, followers, and followees on Instagram, one may obtain, if made available by the user, their age, location, date of account creation, and previously used usernames. On TikTok, the accessible metadata is only the number of followers and followees, and the sum of likes. In some cases, a mixed approach can be problematic to apply, since it can be a time-consuming process containing complex steps of data adaptation to the research questions. Although the aforementioned mixed strategies are based on the combination of four main methods, new mixed approaches can be developed depending on the research needs.

4 Case study: Change4climate

The five strategies to visual data collection discussed in the previous section can be efficiently used for creating a visual dataset from image-based social media platforms. In this section, I attempt to broaden said methods by presenting the corpus data collected for a SOLSTICE project funded by JPI Climate "2O2CM: Overcoming Obstacles and Disincentives to Climate Change Mitigation: A cross-cutting approach by human and social sciences" (https://change4climate.eu/, hereafter, the Project). The Project aims to identify the obstacles to pro-environmental behavior through the lens of anthropology, communication, linguistics, and social psychology. One of the objectives for research in communication includes a collection and analyses of recent multimodal corpora from Instagram, including the posts concerning climate change by opinion leaders (hereafter OLs) and opinion-leading organizations (henceforth OLOs) from three countries: Belgium, France, and Norway. The corpus of posts' visuals, texts, and metadata is obtained by applying a mixed approach. It should be noted that I present a collection of visual data as a part of multimodal corpora from Instagram since image-centricity is a vital concept of multimodal research (Stöckl, 2020).

In the scale of the Project, I define climate change OLs as engaged and competent individuals, who not only draw attention to climate change-related problems but also signal how others should respond or behave to reduce high consumption lifestyles. They can be politicians, media personalities, climate activists, and influencers who publish their messages on climate change mitigation. According to Wang and Li (2016, p. 117), "opinion leaders can also be organizations." In this context, climate change OLOs are organizations, involved in the dissemination of environmental and climate change-related information, drawing particular attention to proposing solutions of how others should react to contribute to climate change mitigation.

To investigate the data obtained from Instagram accounts of climate change OLs and OLOs, I will conduct multimodal content analysis of the posts regarding food, transport, and energy sectors in three countries between 2021 and 2022. To analyze visuals and texts published by climate change OLs and OLOs altogether, I obtained this data from Instagram by applying a mixed approach to data collection. First, the corpus of climate change-related hashtags was obtained with the help of computational tools. This data was further used to identify the most influential Instagram accounts posting about climate change issues. Second, the accounts of OLs and OLOs were manually selected on Instagram according to the par-

ticular criteria described in section 4.2. Finally, the data obtained from accounts and hashtags was filtered according to metadata to provide a more contextualized corpus for the research.

4.1 Corpus of climate change hashtags

The research within social sciences proposes various methods for identifying OLs: from rather qualitative surveys and interviews to more quantitative network analyses (Bamakan, Nurgaliev, & Qu, 2019). More recent methods to identify OLs include various types of network analyses from social media data since it is "easily observable, and researchers do not need to rely on surveys and self-reported measures" (Walter & Brüggemann, 2020, p. 274). Network analyses are constructed with the help of computational tools and determine not only the connections between particular accounts but also potential clusters of OLs. These methods include building connections networks from an existing corpus (Dubois & Gaffney, 2014) and measuring the number of followers and how far their messages travel (Wu, Hofman, Mason, & Watts, 2011).

The main steps for network analysis to identify OLs include (1) keywords selection according to the research topic, (2) collection of the data from social media containing the keywords, (3) network analysis with the help of computational tools, and (4) identification of the clusters of influential accounts and the connections between them. These steps can be adapted to the research objectives and combined with other methods of identifying accounts of OLs (Boatwright, 2022).

Due to the limited time frame of the Project (three years), I adapted these steps of network analysis to our objectives. As the data collected from hashtags can be used to determine influential accounts (Aleahmad, Karisani, Rahgozar, & Oroumchian, 2016), I collected a thematic corpus of posts that included climate change-related hashtags within the countries of the Project. Since it is impossible to collect Instagram posts based on their geographical position only, I manually selected 39 thematic hashtags related to climate change subjects in Dutch, English, French, and Norwegian, e. g., *#klimaatverandering*, *#climateaction*, *#changementclimatique*, *#klimaendringer*, as an entrance point to the data within the geographical area of the research. The data collected from hashtags included 7907 posts with 15 622 images and 798 videos.[4]

The collected data helped us detect some accounts which largely contribute to the dissemination of information regarding climate change. By identifying the users who publish the most with the obtained hashtag-based data, I selected the most influential accounts of environmental organizations in three countries, e. g., *Youth for Climate*, *Fridays for Future*, *Écoconso*. Unfortunately, no influential accounts of individuals within our target locations were identified at this stage. To broaden the corpora and to include the accounts of individuals in it, an additional selection of the accounts of OLs was applied: the accounts of the selected organizations were a starting point for the manual identification of accounts of OLs.[5]

To resume, the data collected from hashtags revealed some drawbacks of this method of data collection from Instagram. For example, the data needed to be additionally sorted after collection, since many posts were found from the accounts located in other areas than our target countries. In addition to this, the results of this collection did not fully correspond to the needs of the Project. Finally, it was also challenging to correspond to the Project's time scale since the data collection started in September 2021 and the tool for data collection was not able to obtain all the posts in the selected time frame due to the large number of posts with certain hashtags. Despite these limitations, the collected hashtag-based data allowed us to identify influential accounts of environmental organizations in Belgium, France, and Norway, which contributed to the further selection of accounts of climate change OLs and OLOs.

4 A detailed collection workflow is introduced in section 4.3.

5 OLs' and OLOs' manual selection process and criteria are presented in the next section.

4.2 Identifying the accounts of opinion leaders and collecting the data

Previously selected accounts from hashtag-based data were used to complement the corpus with the accounts of influential individuals. I manually identified accounts of climate change OLs and OLOs. To do so, a new account on Instagram was created to minimize the influence of the platform's algorithms with previous private accounts. Secondly, I looked at the first 500 accounts of followers of environmental organizations determined as influential accounts from the collected hashtag-based data described in section 4.1. Thirdly, during this observation, I made a separate list of accounts that post about climate change-related topics. These accounts were accounts of both individuals and organizations and had to be open and public to guarantee further data collection. Finally, the selected accounts from step three were further manually evaluated according to five criteria of opinion leadership (Corey, 1971; Flynn, Goldsmith, & Eastman, 1996; Nisbet & Kotcher, 2009; Turcotte, York, Irving, Scholl, & Pingree, 2015; Venkatraman, 1989):

1) *Enduring involvement and information sharing*: to guarantee this, climate change OLs and OLOs should have written about climate change and environment-related topics since January 2021. Users' identification of themselves as involved actors contributed to this criterion. To evaluate this, I suggested that the accounts should mention climate and environment-related words or emoji on their profile description.
2) *Interpersonal influence*: this criterion was measured through the number of followers, which had to be more than 1000: this number results from a developing notion of micro-influencing accounts in marketing (Rios de Castro Marques, Casais, & Camilleri, 2020).
3) *Expertise*: to evaluate the user's level of expertise in environment-related issues, I suggest verifying whether they add references to their information sources. These references can be introduced in visual or verbal modes.
4) *Intermediality*: to measure a user's *intermediality*, I suggest confirming if the account appears on other platforms and websites at least three times. This should not include their profiles on other social media.
5) *Interconnectedness*: to verify the account's *interconnectedness*, I consider mentions in their posts of at least three other accounts related to climate change topics.

The manual evaluation based on five characteristics of opinion leadership allowed us to confirm or disapprove the selected accounts as climate change OLs or OLOs. Overall, 126 accounts constitute the corpus, of which 76 are OLs and 50 are OLOs.

With the help of computational tools,[6] I obtained 13 528 posts by OLs including 20 458 images and 1 949 videos. The corpus of OLOs comprises 9 835 posts with 16 599 images and 1441 videos. After completing the collection of the posts, all data was filtered by date, since the timeframe for the research lies within 2021 and 2022. This allowed us to construct a multimodal corpus,[7] which would help answer our research questions.

4.3 Collecting the data from Instagram

The data collection from Instagram comprised two steps. First, the file with metadata was obtained with *Vurku PRO* services. Second, image files were collected with the help of *ImageDownloader*.

Due to the time restriction of the Project, we (project's investigators from the Belgian team) obtained a PRO subscription to the *Vurku* services. This allowed us an unlimited number of collections of up to 500 most recent Instagram posts from one account, or of posts including a particular hashtag for a very short time (40–60 minutes). Scraping with *Vurku* obtained such posts' metadata as filename, source URL, post date / time, username, followers, engagement rate, like count, comment count, location (if mentioned), post type, caption, hashtags, mentions, image / video URL. The metadata is stored in an Excel file automatically generated by *Vurku*.

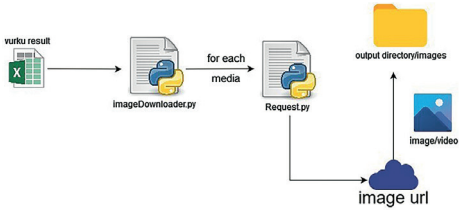
6 The collection process is described in section 4.3.

7 The pseudonymized corpus data is available on the project's website: https://change4climate.eu/toolbox/corpora/

Although other more expensive tools or services can provide an automatic collection of media files, *Vurku* only allows a manual collection of media files. To optimize the process of visual data collection we developed a Python script, that we called *Image Downloader*.[8]

Image Downloader obtains all visual contents from the Excel file created by *Vurku*. Given an Excel file with a column of names and a column of image / video URLs, *Image Downloader* automatically saves the content in the newly created output directory; see Figure 1 for the workflow.

Figure 1: Image Downloader workflow



To sum up, the presented two-step method of collecting posts from Instagram allowed us to build multimodal corpora for the Project. The collection started in September 2021 and ended in March 2022. The limitations and future perspectives of this method are discussed in the following section.

5 Discussion and conclusion

In this paper, I drew our attention to the strategies that can help select and collect visual data from social media for research within communication and media studies. After having carried out a literature review, I synthesized five strategies to visual data collection from image-based social media platforms: hashtag-based, account-based, metadata-based, random sampling, and mixed approaches. These strategies complement the research of Mayr and Weller (2016) and can be effectively implemented for different purposes.

As much research on hashtag-based and account-based data from different social media exists in the communication field, they can be considered basic strategies for social media research. The data collected with these strategies can provide an overview of online communication within the field of interest. They both can be excellent options for longitudinal research. In this case, the research will be influenced by changing platform policies, which is often the case with Instagram and Facebook, and the affordances of the data collection tools. For instance, when Instagram changes access to its API, it can become impossible to obtain data from the platform. Random sampling is one of the most accessible strategies since a variety of corpora can be provided by other institutions. Nevertheless, random samples more often provide data on multiple subjects or by multiple communicators: In some cases, it may be complicated to achieve sufficient levels of corpus homogeneity and representativity. The metadata-based strategy is more frequently implied in combination with other strategies and helps to narrow the dataset down or contextualize the research. Mixed approach combines multiple strategies in order to contextualize the corpus and can be more suitable for a qualitative research. All exemplary studies provided in section 3.5 combined two strategies to provide a precise dataset. However, depending on the research question, three and more strategies can represent a mixed approach, if the time and tool resources are sufficient for this purpose. All exemplary studies of mixed approach in section 3.5 are rather related to post-collection data processing than to data collection itself. However, some tools to collect the data from Instagram and YouTube provide a filtering function that can be used during the data collection. For example, *Instaloader* (https://instaloader.github.io/) and *youtube-dl* (https://github.com/ytdl-org/youtube-dl) can provide a collection of posts or videos from one account within a specific timeframe or with a specific minimum or maximum number of views.

Various quantitative and qualitative methods in visual communication research are applied to the present strategies, such as content analysis or framing analysis. Since the

8 This code is free to use and can be obtained on request.

strategies themselves can influence the selected content of the sample and, therefore, change the results of any subsequent analysis (McKitrick, Schuurman, & Crooks, 2023), particular attention should be made to the choice of the collection methods at the preliminary stage of the research. It is also important to note that all presented strategies for data collection are related to the analysis of the content itself, without attempting to reach the people behind the screen. More research is needed considering the combination of the analyses of social media visual content with real-time observations or interviews.

According to our observations, the presented strategies are related to a classic approach to social media data collection, where text functions as context to the visual stimuli. In the case where only the visual component is taken for the analysis, it might present a significant limitation to data selection. More precisely, each hashtag of the post is a verbal component, which is extracted from a post's text. Depending on the context, the image might or might not be related to this theme. Therefore, more accurate methods and tools to access the visuals from social media by not considering the textual component should be developed.

To illustrate a mixed approach to visual data collection from Instagram, I provided a case study involving a hashtag-account-metadata mixed approach. As shown by the study of Walter and Brüggemann (2020) on Twitter data about COP21 meetings, hashtag-based data can be efficiently used to obtain a specific thematic corpus to identify OLs. In our case, however, this method was not sufficient to provide an extensive corpus of messages by climate change OLs and OLOs on Instagram. A possible reason for this could be the particular use of hashtags by environmental organizations. This hypothesis should be studied in detail by further research. To complement the corpus of OLOs from hashtag-based data, I manually selected accounts of climate change OLs. Metadata-based filtering contributed to the contextualization of a dataset for the analysis regarding the Project's needs: timeframe between 2021 and 2022.

Due to the research aim, our data required multi-level processing. Saving the visuals and metadata in adequate formats by using computational tools was a significant part of data collection. After collecting the data, pseudonymization is implemented to the accounts of OLs to ensure the ethical requirements. During the analysis phase, automated text tagging is applied to select a subcorpus of posts about climate change mitigation in food, energy, and transport sectors, and a qualitative multimodal content analysis is applied to closely investigate the posts.

The presented mixed approach to Instagram data collection and sorting can be used to work with visual data shared on this social network, although it requires some improvements. First of all, the data collection from Instagram for research purposes needs to be a more rapid process, accessible for users with different levels of programming experience. In addition to this, it is necessary to develop accessible and reliable open-source data collection tools from image-based social media, which are also based on accessing the data from visual and multimodal components of the posts. While I tried to minimize the algorithms bias of Instagram by creating a new account, more detailed research should be done to investigate the algorithmic bias in the data selection process. Synthesized information about the algorithms of all social media platforms could be an essential document for the researchers in digital communication field.

To conclude, studying new forms of visual communication on social media is an interesting and challenging process. Due to the variety of social media platforms and a large number of posts about any imagined topic shared every second in the whole world, it can be complicated to choose the right data for research. In this article, I attempted to systematize the most recent approaches to visual – as part of multimodal – data collection process from image-based social media. The exemplary studies presented in section 3 demonstrated the variety of strategies for visual data collection and can serve as a guide for future research. The case study demonstrated that identification of influential accounts from hashtag-based data does not

always provide the necessary results, but can contribute to further manual selection of accounts of opinion leaders and opinion-leading organizations. Furthermore, filtering based on metadata can help contextualize the research data. Finally, I suggest that the presented scheme can be replicated to identify influential accounts in other spheres, such as politics, art, or health. Future research in communication and media studies could address other types of mixed data collection involving multiple image-based social media.

Acknowledgements

I would like to thank all people involved in the project, in particular my supervisors Prof Andrea Catellani and Dr Louise-Amélie Cougnon for their guidance and support. This work is funded by the JPI Climate and BELSPO as a part of the 2O2CM project (http://change4climate.eu).

Conflict of interests

The author declares no conflict of interest.

References

Abdesslem, F. B., Parris, I., & Henderson, T. (2012). Reliable online social network data collection. In A. Abraham (Ed.), *Computational social networks* (pp. 183–210). London, UK: Springer. https://doi.org/10.1007/978-1-4471-4054-2_8

Ahrens, J., Brennan, F., Eaglesham, S., Buelo, A., Laird, Y., Manner, J., … Sharpe, H. (2022). A longitudinal and comparative content analysis of Instagram fitness posts. *International Journal of Environmental Research and Public Health*, *19*(11), 1–13. https://doi.org/10.3390/ijerph19116845

Aleahmad, A., Karisani, P., Rahgozar, M., & Oroumchian, F. (2016). OLFinder: Finding opinion leaders in online social networks. *Journal of Information Science*, *42*(5), 659–674. https://doi.org/10.1177/0165551515605217

Al-khateeb, S., & Agarwal, N. (2019). Tools and methodologies for data collection, analysis, and visualization. In S. Al-khateeb & N. Agarwal (Eds.), *Deviance in social media and social cyber forensics: Uncovering hidden relations using open source information (OSINF)* (pp. 45–65). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-13690-1_3

Allgaier, J. (2019). Science and environmental communication on YouTube: Strategically distorted communications in online videos on climate change and climate engineering. *Frontiers in Communication*, *4*, 1–15. https://doi.org/10.3389/fcomm.2019.00036

Aslam, S. (2023, February 28). Instagram by the numbers: Stats, demographics & fun facts [Website]. Retrieved from https://www.omnicoreagency.com/instagram-statistics/

Bamakan, S. M. H., Nurgaliev, I., & Qu, Q. (2019). Opinion leader detection: A methodological review. *Expert Systems with Applications*, *115*, 200–222. https://doi.org/10.1016/j.eswa.2018.07.069

Barthes, R. (1967). *Elements of semiology* (A. Lavers & C. Smith, Trans.). New York, NY: Hill and Wang.

Boatwright, B. C. (2022). Exploring online opinion leadership in the network paradigm: An analysis of influential users on Twitter shaping conversations around anthem protests by prominent athletes. *Public Relations Review*, *48*(4), 1–12. https://doi.org/10.1016/j.pubrev.2022.102229

Boling, K. S. (2020). #ShePersisted, Mitch: A memetic critical discourse analysis on an attempted Instagram feminist revolution. *Feminist Media Studies*, *20*(7), 966–982. https://doi.org/10.1080/14680777.2019.1620821

Bossetta, M. (2018). The digital architectures of social media: Comparing political campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U. S. election. *Journalism & Mass Communication Quarterly*, *95*(2), 471–496. https://doi.org/10.1177/1077699018763307

Corey, L. G. (1971). People who claim to be opinion leaders: Identifying their characteristics by self-report. *Journal of Marketing*, *35*(4), 48–53. https://doi.org/10.1177/002224297103500409

Cougnon, L.-A., de Viron, L., & Watrin, P. (2022, January 29). *Collection of Twitter corpora for human and social sciences: Sampling methodology and evaluation*. Presented at

the LREC conference 2022, 13th Edition of its Language Resources and Evaluation Conference of the ELRA. Marseille, France. https://doi.org/10.31235/osf.io/puhw8

Dubois, E., & Gaffney, D. (2014). The multiple facets of influence: Identifying political influentials and opinion leaders on Twitter. *American Behavioral Scientist*, *58*(10), 1260–1277. https://doi.org/10.1177/0002764214527088

Ferchaud, A., Grzeslo, J., Orme, S., & LaGroue, J. (2018). Parasocial attributes and YouTube personalities: Exploring content trends across the most subscribed YouTube channels. *Computers in Human Behavior*, *80*, 88–96. https://doi.org/10.1016/j.chb.2017.10.041

Flynn, L. R., Goldsmith, R. E., & Eastman, J. K. (1996). Opinion leaders and opinion seekers: Two new measurement scales. *Journal of the Academy of Marketing Science*, *24*(2), 137–147. https://doi.org/10.1177/0092070396242004

Fung, I. C.-H., Blankenship, E. B., Ahweyevu, J. O., Cooper, L. K., Duke, C. H., Carswell, S. L., … Tse, Z. T. H. (2019). Public health implications of image-based social media: A systematic review of Instagram, Pinterest, Tumblr, and Flickr. *The Permanente Journal*, *24*(1), 1–10. https://doi.org/10.7812/TPP/18.307

Gibbs, M., Meese, J., Arnold, M., Nansen, B., & Carter, M. (2015). #Funeral and Instagram: Death, social media, and platform vernacular. *Information, Communication & Society*, *18*(3), 255–268. https://doi.org/10.1080/1369118X.2014.987152

Gruzd, A., Lannigan, J., & Quigley, K. (2018). Examining government cross-platform engagement in social media: Instagram vs Twitter and the big lift project. *Government Information Quarterly*, *35*(4), 579–587. https://doi.org/10.1016/j.giq.2018.09.005

Hartika, M., Pawito, & Utari, P. (2022). Brand activism on the digital public sphere: Campaign content analysis of #BringBackOurBottle on Instagram. *IOP Conference Series: Earth and Environmental Science*, *1016*(1), 1–7. https://doi.org/10.1088/1755-1315/1016/1/012027

Hassan, S. Z., Ahmad, K., Hicks, S., Halvorsen, P., Al-Fuqaha, A., Conci, N., & Riegler, M. (2022). Visual sentiment analysis from disaster images in social media. *Sensors*, *22*(10), 1–21. https://doi.org/10.3390/s22103628

Hautea, S., Parks, P., Takahashi, B., & Zeng, J. (2021). Showing they care (or don't): Affective publics and ambivalent climate activism on TikTok. *Social Media + Society*, *7*(2), 1–14. https://doi.org/10.1177/20563051211012344

Highfield, T., & Leaver, T. (2016). Instagrammatics and digital methods: Studying visual social media, from selfies and GIFs to memes and emoji. *Communication Research and Practice*, *2*(1), 47–62. https://doi.org/10.1080/22041451.2016.1155332

Kim, H., Jang, S. M., Kim, S.-H., & Wan, A. (2018). Evaluating sampling methods for content analysis of Twitter data. *Social Media + Society*, *4*(2), 1–10. https://doi.org/10.1177/2056305118772836

Kim, Y., Song, D., & Lee, Y. J. (2020). #Antivaccination on Instagram: A computational analysis of hashtag activism through photos and public responses. *International Journal of Environmental Research and Public Health*, *17*(20), 1–20. https://doi.org/10.3390/ijerph17207550

Leaver, T., Highfield, T., & Abidin, C. (2020). *Instagram: Visual social media cultures*. Cambridge, UK: Polity Press.

Lee, E., & Weder, F. (2021). Framing sustainable fashion concepts on social media. An analysis of #slowfashionaustralia Instagram posts and post-COVID visions of the future. *Sustainability*, *13*(17), 1–17. https://doi.org/10.3390/su13179976

Lee, Y., Huang, K.-T. (Tim), Blom, R., Schriner, R., & Ciccarelli, C. A. (2021). To believe or not to believe: Framing analysis of content and audience response of top 10 deepfake videos on YouTube. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 153–158. https://doi.org/10.1089/cyber.2020.0176

León, B., Negredo, S., & Erviti, M. C. (2022). Social engagement with climate change: Principles for effective visual representation on social media. *Climate Policy*, *22*(8), 976–992. https://doi.org/10.1080/14693062.2022.2077292

Li, Y., & Xie, Y. (2020). Is a picture worth a thousand words? An empirical study of image content and social media engagement. *Journal of Marketing Research*, *57*(1), 1–19. https://doi.org/10.1177/0022243719881113

Lindgren, S. (2018). The concept of "data" in digital research. In U. Flick (Ed.), *The SAGE handbook of qualitative data collection* (pp. 441–450). London, UK: Sage.

Loizos, P. (2000). Video, film and photographs as research documents. In M. W. Bauer & G. Gaskell (Eds.), *Qualitative researching with text, image and sound* (pp. 94–107). London, UK: Sage. https://doi.org/10.4135/9781849209731.n6

Mapes, G., & Ross, A. S. (2022). Making privilege palatable: Normative sustainability in chefs' Instagram discourse. *Language in Society*, *51*(2), 259–283. https://doi.org/10.1017/S0047404520000895

Mayr, P., & Weller, K. (2016). Think before you collect: Setting up a data collection approach for social media studies. In A. Quan-Haase & L. Sloan (Eds.), *The SAGE handbook of social media research methods* (pp. 107–124). London, UK: Sage.

McKitrick, M. K., Schuurman, N., & Crooks, V. A. (2023). Collecting, analyzing, and visualizing location-based social media data: Review of methods in GIS-social media analysis. *GeoJournal*, *88*, 1035–1057. https://doi.org/10.1007/s10708-022-10584-w

Moffitt, B. (2024). How do populists visually represent "The People"? A systematic comparative visual content analysis of Donald Trump and Bernie Sanders' Instagram accounts. *The International Journal of Press/Politics, 29*(1), 74–99. https://doi.org/10.1177/19401612221100418

Molder, A., Chen, K., Clemmons, Z., & Lakind, A. (2021). Framing the global youth climate movement: A qualitative content analysis of Greta Thunberg's moral, hopeful, and motivational framing on Instagram. *The International Journal of Press/Politics*, *27*(3), 668–695. https://doi.org/10.1177/19401612211055691

Nau, C., Quan-Haase, A., & McCay-Peet, L. (2022). Defining social media and asking social media research questions: How well applies the swiss army knife metaphor? In A. Quan-Haase & L. Sloan (Eds.), *The SAGE handbook of social media research methods* (2nd ed., pp. 13–26). London, UK: Sage.

Nisbet, M. C., & Kotcher, J. E. (2009). A two-step flow of influence? Opinion-leader campaigns on climate change. *Science Communication*, *30*(3), 328–354. https://doi.org/10.1177/1075547008328797

O'Halloran, K., Chua, A., & Podlasov, A. (2014). The role of images in social media analytics: A multimodal digital humanities approach. In D. Machin (Ed.), *Visual communication* (pp. 565–587). Berlin, Germany: De Gruyter Mouton.

Pearce, W., Niederer, S., Özkula, S. M., & Sánchez Querubín, N. (2019). The social media life of climate change: Platforms, publics, and future imaginaries. *WIREs Climate Change*, *10*(2), 1–13. https://doi.org/10.1002/wcc.569

Pearce, W., Özkula, S. M., Greene, A. K., Teeling, L., Bansard, J. S., Omena, J. J., & Teixeira Rabello, E. (2020). Visual cross-platform analysis: Digital methods to research social media images. *Information, Communication & Society*, *23*(2), 161–180. https://doi.org/10.1080/1369118X.2018.1486871

Pramana, P. D., Utari, P., & Naini, A. M. I. (2021). Symbolic convergence of #ClimateCrisis: A content analysis of Greenpeace Indonesia campaign on Instagram. *IOP Conference Series: Earth and Environmental Science*, *724*(1), 1–7. https://doi.org/10.1088/1755-1315/724/1/012101

Rieder, B., Matamoros-Fernández, A., & Coromina, Ò. (2018). From ranking algorithms to "ranking cultures": Investigating the modulation of visibility in YouTube search results. *Convergence*, *24*(1), 50–68. https://doi.org/10.1177/1354856517736982

Rios de Castro Marques, I., Casais, B. G., & Camilleri, M. A. (2020). The effect of macro celebrity and micro influencer endorsements on consumer-brand engagement on Instagram. In M. A. Camilleri (Ed.), *Strategic corporate communication in the digital age* (pp. 1–20). Bingley, UK: Emerald. Retrieved from https://papers.ssrn.com/abstract=3705334

Rogers, R. (2017). Digital methods for cross-platform analysis. In J. Burgess, A. Marwick, & T. Poell (Eds.), *The SAGE handbook of social media* (pp. 91–110). London, UK: Sage.

Rutherford, B. N., Sun, T., Johnson, B., Co, S., Lim, T. L., Lim, C. C. W., … Chan, G. C. K. (2022). Getting high for likes: Exploring cannabis-related content on TikTok. *Drug and Alcohol Review*, *41*(5), 1119–1125. https://doi.org/10.1111/dar.13433

Samani, Z. R., Guntuku, S. C., Moghaddam, M. E., Preoţiuc-Pietro, D., & Ungar, L. H. (2018). Cross-platform and cross-interaction study of user personality based on images on Twitter and Flickr. *PLOS ONE*, *13*(7), 1–19. https://doi.org/10.1371/journal.pone.0198660

Schroeder, R. (2018). *Social theory after the Internet: Media, technology, and globalization*. London, UK: UCL Press.

Shi, C.-F., So, M. C., Stelmach, S., Earn, A., Earn, D. J. D., & Dushoff, J. (2022). From science to politics: COVID-19 information fatigue on YouTube. *BMC Public Health*, *22*, 1–14. https://doi.org/10.1186/s12889-022-13151-7

Smith, A. N., Fischer, E., & Yongjian, C. (2012). How does brand-related user-generated content differ across YouTube, Facebook, and Twitter? *Journal of Interactive Marketing*, *26*(2), 102–113. https://doi.org/10.1016/j.intmar.2012.01.002

Statista. (2022, April). Media usage in an Internet minute as of April 2022 [Website]. Retrieved from https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*, 156–168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002

Stöckl, H. (2020). Multimodality and mediality in an image-centric semiosphere – A rationale. In C. Thurlow, C. Dürscheid, & F. Diémoz (Eds.), *Visualizing digital discourse: Interactional, institutional and ideological perspectives* (pp. 189–202). Berlin, Germany: De Gruyter Mouton. https://doi.org/10.1515/9781501510113-010

Su, L. Y.-F., McKasy, M., Cacciatore, M. A., Yeo, S. K., DeGrauw, A. R., & Zhang, J. S. (2022). Generating science buzz: An examination of multidimensional engagement with humorous scientific messages on Twitter and Instagram. *Science Communication*, *44*(1), 30–59. https://doi.org/10.1177/10755470211063902

Tifentale, A., & Manovich, L. (2015). Selfiecity: Exploring photography and self-fashioning in social media. In D. M. Berry & M. Dieter (Eds.), *Postdigital aesthetics: Art, computation and design* (pp. 109–122). London, UK: Palgrave Macmillan. https://doi.org/10.1057/9781137437204_9

Trillò, T., Scharlach, R., Hallinan, B., Kim, B., Mizoroki, S., Frosh, P., & Shifman, L. (2021). What does #freedom look like? Instagram and the visual imagination of values. *Journal of Communication*, *71*(6), 875–897. https://doi.org/10.1093/joc/jqab021

Turcotte, J., York, C., Irving, J., Scholl, R. M., & Pingree, R. J. (2015). News recommendations from social media opinion leaders: Effects on media trust and information seeking. *Journal of Computer-Mediated Communication*, *20*(5), 520–535. https://doi.org/10.1111/jcc4.12127

Veltri, G. A., & Atanasova, D. (2017). Climate change on Twitter: Content, media ecology and information sharing behaviour. *Public Understanding of Science*, *26*(6), 721–737. https://doi.org/10.1177/0963662515613702

Venkatraman, M. P. (1989). Opinion leaders, adopters, and communicative adopters: A role analysis. *Psychology & Marketing*, *6*(1), 51–68. https://doi.org/10.1002/mar.4220060104

Veum, A., & Undrum, L. V. M. (2018). The selfie as a global discourse. *Discourse & Society*, *29*(1), 86–103. https://doi.org/10.1177/0957926517725979

Walter, S., & Brüggemann, M. (2020). Opportunity makes opinion leaders: Analyzing the role of first-hand information in opinion leadership in social media networks. *Information, Communication & Society*, *23*(2), 267–287. https://doi.org/10.1080/1369118X.2018.1500622

Wang, Y., & Li, Y. (2016). Proactive engagement of opinion leaders and organization advocates on social networking sites. *International Journal of Strategic Communication*, *10*(2), 115–132. https://doi.org/10.1080/1553118X.2016.1144605

Waters, R. D., & Jones, P. M. (2011). Using video to build an organization's identity and brand: A content analysis of nonprofit organizations' YouTube videos. *Journal of Nonprofit & Public Sector Marketing*, *23*(3), 248–268. https://doi.org/10.1080/10495142.2011.594779

Weller, W., Bassalo, L. M. B., & Pfaff, N. (2018). Collecting data for analyzing blogs. In U. Flick (Ed.), *The SAGE handbook of qualitative data collection* (pp. 482–495). London, UK: Sage. https://doi.org/10.4135/9781526416070

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Who says what to whom on Twitter. *WWW '11: Proceedings of the 20th International Conference on World Wide Web*, 705–714. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/1963405.1963504

Zeng, J., & Abidin, C. (2021). "#OkBoomer, time to meet the Zoomers": Studying the memefication of intergenerational politics on TikTok. *Information, Communication & Society*, *24*(16), 2459–2481. https://doi.org/10.1080/1369118X.2021.1961007