Accelerating Research with Multi-National, Multi-Platform Image Archives

Cody Buntain

16 June 2022

Abstract

This report outlines a set of scientific challenges around the use and analysis of images in online political discourse and related computational social science research tasks. We identify four main gaps: 1) lack of sufficient automated methods for characterizing the variety and complexity of images used in political discussion, and 2) the absence of baselines describing "normal" behavior by online audiences – thereby complicating our understanding what is abnormal, especially among influence campaigns. Relatedly, 3) demonstrates a gap in that much of the current frameworks for image analysis do not clearly support search methods for finding accounts sharing relevant similar imagery. Finally, (4) examines the extant barriers to using this data. Reducing these barriers could then allow expertise from computational social science scholars to mitigate some of the more blatant biases that purely computational approaches might produce. After outlining these gaps, we then propose a potential infrastructure and framework for accelerating this work.

Introduction

When asked about open questions concerning studies of influence efforts, a common refrain is to bemoan how little we know about the role of visual media in online influence. This gap stems from a confluence of factors, including the relative difficulty in analyzing image data, the relative scarcity of image-oriented datasets around online political discourse, and the significant storage infrastructure needed when working with large quantities of images as compared to textual data. This latter issue is easily illustrated by an examination of Twitter's Information Operations datasets,[1] where Twitter has provided both the text and image data from accounts associated with malevolent influence campaigns: As of March 2022, this archive is approximately 9.5 terabytes in size, 99.7% of which is comprised of media files (i.e., only about 30 gigabytes of this archive are text data). This imbalance between text and image data is not specific to influence efforts either, as demonstrated by a collection of crisis-related social media content across 100 crisis events, where textual data accounts for 3% of a 40-gigabyte archive, with image data accounting for the other 97%.[2]

While challenges related to dataset sizes (in terms of bytes needed) can be mitigated with a sufficiently well-resourced storage infrastructure, and large-scale image-processing issues can be reduced with sufficiently sizable high-performance or cloud-computing infrastructure, simply increasing scale does not address core scientific issues:

- The research community generally lacks access to automated methods capable of characterizing the political or discursive context of visual media in online spaces. Though computer-vision and image-understanding methods are maturing rapidly, how applicable these methods are to political discourse and their robustness to the variety of visual media shared (e.g., photographs versus drawings versus image macros) remains an open question.
- The research community also has limited access to consistent baselines of discourse and media use in the online environment, an omission that hinders development and evaluation in automated visual analytics methods – that is, without a good sample of "normal" behavior in visual media sharing against which we can compare, understanding what is different or unique about a particular group's use of media becomes difficult.
- Limited infrastructure exists for identifying and tracking what accounts are sharing and amplifying particular instances of visual media. Whereas textual modalities like hashtags, websites, or usernames are easily identifiable, identifying which accounts have amplified a particular images is more difficult, as few frameworks or API offerings exist for image-based

[1] Twitter Transparency, "Information Operations," Twitter, 2022, https://transparency.twitter.com/en/reports/information-operations.html.

[2] TREC Incident Streams, "TREC Incident Streams Track Homepage," University of Glasgow, 2020, http://trecis.org.

search in online social platforms (though certain services, like CrowdTangle, have offered such a capability). This topic in particular has been highlighted across numerous interviews conducted for this report.

- Lastly, much of the expertise in the image processing and computer vision techniques used for characterizing visual media is siloed away from social science and computational social science scholars. Though the techniques and pre-trained models exist to support analytical frameworks similar to modern text analysis, issues such as data access and scalability of processing and storage have been major barriers for these scholars. Consequently, much of the work on image processing has focused on object recognition and search tasks and omit key questions around images in political discourse, hate speech, etc. Reducing these barriers to access could then allow expertise from computational social science scholars to mitigate some of the more blatant biases that purely computational approaches might produce.

All these issues are then compounded by the speed at which these online platforms evolve and the increasing centrality of media in these spaces, which has implications for the temporal validity of any findings associated with these spaces.

This report outlines research challenges and proposes the blueprints for an infrastructure to address these issues. This blueprint describes the creation of a standardized collection of images along with the infrastructure to continuously update and maintain these test collections, reduce barriers for social science scholars, and accelerate their research. We envision this potential resource as containing a widely available collection of datasets and a sample of accounts plus the visual media they share, selected from a variety of national contexts and online social platforms. This test collection could then provide the data necessary to enhance validity and evaluate what truly is different in the behavior of a particular online group compared to the general audience. Additionally, making these datasets available along with pre-processed versions of their contents could drastically reduce barriers to entry for computational social science scholars and scholars from less resourced institutions or countries. This availability of pre-processed images (i.e., their dense embeddings) could both alleviate scalability issues by offloading and centralizing a computationally intensive step and providing a research pipeline analogous to modern word-embedding- and neural language model-based text analyses.

Crucially, these collections should not exist solely as isolated, static resources; instead, the ongoing collection of this data could further enable researchers to identify cascades of image sharing and identify which accounts engage in the propagation of a particular image, an especially important phenomenon now as seen in the propagation of visuals from the Ukraine-Russia war. The infrastructure to maintain and continuously expand this test collection could also allow researchers to

address questions of temporal validity in that one can use this data to understand how these online audiences use of media evolve over time.

Making this collection available, consolidating the infrastructure needed to maintain this collection, and supporting the execution of common social science research tasks on it represents a substantial opportunity to impact computational social science research. The transformative potential of such a resource to accelerate studies of images in online political conversation and make such work more tractable and accessible for the wider research community is difficult to understate.

Motivations from Image-Oriented Data Challenges and Interviews

Recently, the authors co-organized a workshop on images in political discourse, PhoMemes,[3] at the annual International Conference on the Web and Social Media. This workshop has included a multi-task data challenge, with three tasks:

1. Hateful Imagery Detection – Participants are asked to classify images as containing anti-social symbols or presenting an anti-social message.
2. Identification of Disinformation Agents via their Media – This challenge asks for systems to use image-characterization methods to classify social media accounts into authentic or campaign accounts, where the campaign label is decomposed into sub-labels across several influence campaigns.
3. Attribution of Screenshots to Online Social Platforms – As screenshots are increasingly used for reporting quotes and actions of elites across platforms, methods for tracking and attributing these screenshots to individuals are increasingly useful for establishing provenance and propagation paths. This challenge will identify and cluster duplicate screenshots and ascribe them to their source social media platforms.

To populate the data challenges, we made use of Twitter's Information Operations datasets, where we first encountered scalability issues around the sizes of image archives. For example, a dataset of approximately 50,000 tweets occupies several gigabytes of compressed storage, and an increase in order of magnitude (e.g., 50,000 to 500,000 images) similarly increases the size requirement for such content by the same order. Hence, trying to make an archive of images broadly available has required a substantial downsampling of accounts and images per account. The proposed infrastructure outlined herein would provide an invaluable resource for sharing test collections for community challenges.

Over this workshop and related keynotes, two additional motivations for expanding image analysis in online discourse became apparent. First, as outlined in the keynote by Sefa Ozalp, principal data

[3] PhoMemes 2022, "PhoMemes 2022," PhoMemes, 2022, https://phomemes.github.io/.

scientist at the ADL's Center on Extremism, if one were to ignore images in assessing online discourse, one would miss more than 50% of relevant conversation. These images carry significant signal about context in the discussion. At the same time, as discussed by the second keynote, Prof. Kevin Munger, and several authors in the workshop, understanding the meaning of these images – and memes in particular – is challenging even for manual assessment, suggesting a substantial opportunity for advancing the state of the art in this space. A major point raised here was the need for methods that looked across multiple modalities, joining the visual, textual, and social, to capture meaning in these multimodal conversations. Likewise, several authors echoed needs for standardized datasets and image-search frameworks.

In the lead up to and aftermath of this workshop, we also circulated a survey, asking researchers to share their thoughts on how one might accelerate image analysis for computational social science. Core questions in this survey were:

- How much data is sufficient for you?
- How do you deal with scale in storage?
- How do you currently handle image characterization?
- How do you search the images you have?
- What are the weaknesses in your image analysis pipelines?
- What would accelerate your research the most?

In discussions with Muhammad Imran of QCRI (creator for the CrisisMMD multimodal dataset for crisis informatics), Sefa Ozalp, and others, common answers to these questions were:

- Thousands to millions of images are needed for high-quality analysis.
    - Prof. Imran in particular suggested at least three million images were needed to perform reasonable automated analyses.
    - This scale roughly translates to hundreds of gigabytes of image data.
- Scaling issues were generally handled piecemeal and with a combination of university and cloud computing resources.
- Image characterization primarily used off-the-shelf, pre-trained models, particularly ResNet50, with others using variations on deep neural networks.
- Despite consistent interest and desire, few respondents had answers for searching their datasets of images, ranging from ad-hoc queries to using Google Images API to manual assessment.
- Primary weaknesses focused on scalability issues and scarcity of good datasets.
- Relatedly, common requests about accelerating research focused on availability of pre-built datasets, access to image-search infrastructure, and better models for characterizing the various types of imagery in political discourse.

Taken together, these responses outline critical weaknesses and present significant opportunities for enabling transformative research in this space.

Proposed Archive Structure and Recommendations

Figure 1 shows a high-level view of a possible architecture for such an archive. The data collection framework could sample accounts according to given criteria, such as the account's political interest or geolocated national context. This framework is meant to be extensible to support image collection across multiple platforms. Raw images from these platforms and the text associated with the message from which the image is extracted can then flow into the image storage component, which uses existing image characterization models to generate dense embeddings for a sample of each account's stored images. This dense embedding captures some aspects of the associated image's content, such that two images with similar embeddings should be visually similar.

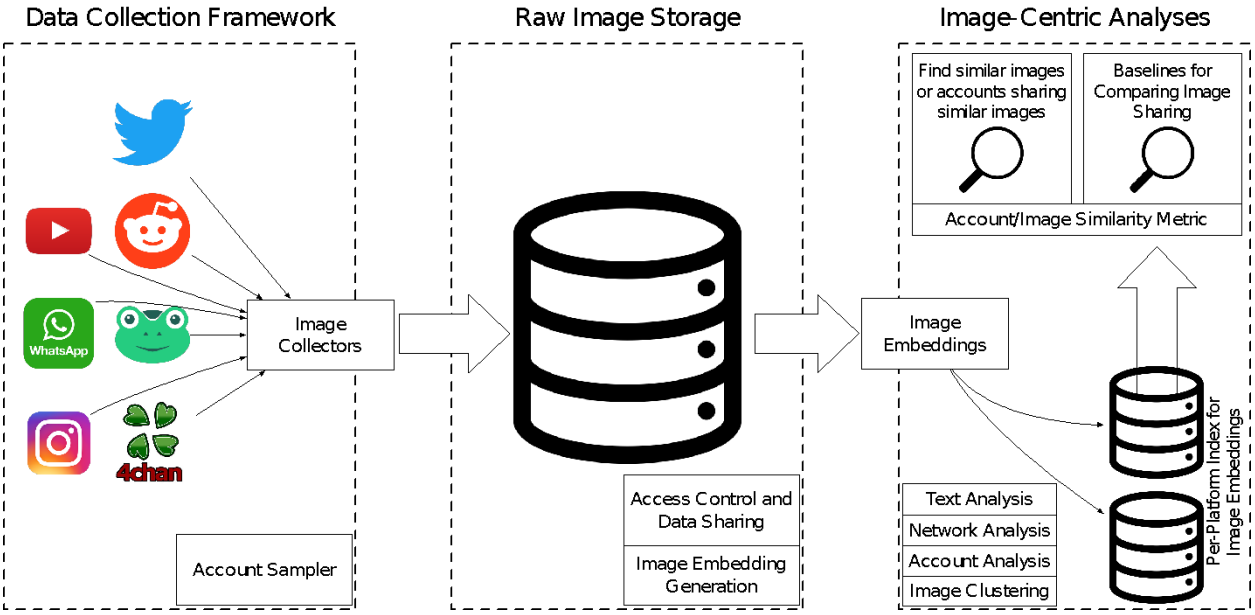

Figure 1. Image Archive and Characterization Infrastructure

These embeddings could then flow into an analytics module, which indexes source-account information, associated text, and the image's embeddings, thereby reducing the memory footprint for each index. We envision additional simple analyses running on top of these image embeddings to provide account-level aggregation and image-type characterization via clustering. From this module, researchers can access the baseline collection of accounts from a given national and/or political context for comparison, and researchers could search for accounts sharing images similar to a given query image to study propagation – a key research task currently lacking in the vast majority of image frameworks.

Architecture Platform Selection

Initially, one could target Twitter and Reddit as primary accounts, as data collection from these spaces is relatively straightforward. Alt-tech platforms like Gab could be added as well. Video-oriented platforms like YouTube, TikTok, BitChute, and Rumble could be included in follow-on work by identifying key images in a given video (e.g., thumbnail identification as a first pass).

Account Selection

Ideally, one might target sets of 10,000 active accounts (where "active" is defined as sharing more than some threshold of messages per month), in two contexts: general audience and political audience. For general audiences, one could sample user IDs or screen names via hashing to construct a random sample of accounts that engaged in particular conversation or shared content associated with particular/national context. For political samples, in Twitter, one can identify accounts who follow a collection of known political elites. For Reddit, one similarly could identify accounts that engage in known politically relevant subreddits/communities. In both cases, however, one should maintain and update these lists of political elites and communities over time.

Data Collection

For accounts sampled, this infrastructure could collect and store all the images shared in a public setting by a sample of relevant accounts. These images could then be stored in raw form on disk, to support retrospective and qualitative analysis of the images shared. We anticipate the distribution of images shared per account to be log-normal, with few accounts sharing many images and the majority of accounts sharing a few images.

To address concerns associated with the scale of these collections, interviews suggest successful methods for storing these collections use object-stores such as Amazon's Simple Storage Service (S3), available as part of Amazon Web Services (AWS).

Scaling Image Characterization and Supporting Search

While the proposed infrastructure stores the raw form of these images, we also recommend the creation of a database of dense embeddings for each image associated with accounts, to allow rapid analysis and search for similar images. These embeddings will allow researchers to evaluate similarity between images while maintaining small signatures for these images in this dense feature space. Supporting search in this lower-dimensional space has the benefit of a significant reduction in storage footprint for image analysis. Additionally, these dense embeddings are similar to those developed for language modeling. Consequently, researchers already proficient in or familiar with text-based embeddings will be well-positioned to use this data. Furthermore, providing images and embeddings in both contexts allows for a consistent pipeline for understanding multimodal (i.e., both text and

image) content shared by accounts further by making these embeddings available to the research community. This dataset should allow the research community to analyze and characterize these images much more rapidly and without the need for significant infrastructure to generate different betting these different settings on their own. While embeddings using pretraining models are likely to be insufficient for capturing the full political context of an image, these methods allow for a first pass that provides rapid image analysis tools to the wider research community. Additionally given the relatively small storage footprint of the beddings, this framework could store multiple versions of these embedding using different analysis models.

Use Cases

Authentic versus Influence Efforts

While Twitter's electoral integrity datasets provide large volumes of images attributed to influence campaigns, the research community lacks a paired dataset of negative samples. Without these negative instances, it is difficult to evaluate the degree to which authentic accounts are different from members of influence campaigns. Using the datasets and infrastructure outlined herein, one could compare images from the "normal" population samples to those inauthentic disinformation accounts identified by Twitter. These comparisons are crucial for contextualizing analyses of these inauthentic accounts and ensuring analytical results are specific to anti-social groups rather than social media users more broadly.

Quantifying Influence

One approach for assessing the impact of an influence campaign is to evaluate whether discourse within general and political audiences is becoming more or less similar to discourse from those groups engaged in the influence campaign. The infrastructure outlined herein can enable this assessment by measuring whether and which images are more successful in propagating across online audiences. In particular, one could identify images shared by particular accounts and evaluate whether these images are shared by the general population after exposure and whether shared images are becoming more or less similar to those shared by target influence campaigns.

Consistency Across Modalities

While significant work has gone into characterizing the textual behaviors of online actors and influence agents, relatively less work has gone into evaluating behaviors of these accounts and non-textual modalities. Using the resources outlined herein could provide one to evaluate consistency of an account across modalities by looking both at the text they share and at their imagery to evaluate whether they present a consistent narrative or ideological view of themselves across them. One might expect that camouflaging one's self using text is easier than obfuscating one's message in imagery, as

imagery is more expensive to create. That is, given the difficulty or relatively higher costs in creating new media, it might be more difficult for inauthentic agents to appear consistent when sharing images instead of text.

Image Propagation and Screenshots

Current infrastructure does not allow researchers to evaluate or identify accounts that are sharing a particular image with slight modifications of a particular image. This omission makes answering questions such as who is sharing a particular meme or how a particular image is being propagated throughout the patient space difficult to answer. As an example, a recent phenomenon in online social spaces is the increasing use of screenshots as a means to share news information, report first-person observations, and quote someone. As people rely increasingly on sharing screenshots of other people's behavior, the boundaries between platforms become increasingly fuzzy as screenshots of one platform are easily shared as images on a different length. Despite this new behavior, ascribing images of screenshots to source platforms remains a difficult task (one which we have included in the recent PhoMemes workshop and data challenge).

Recommendations on Bringing Text-Analysis Methods to Images

A key value-add to the infrastructure described above is the availability of image-level embeddings. In principle, these dense representations of images can be aggregated up to the account level or clustered into various types of images. More interestingly, however, is the convergence between visual embedding models and recent trends toward text embedding and neural language models. Restated, as neural language models have become more popular, so too have transformers and embedding representations of text – e.g., from the 2013 Word2Vec model to the more recent 2018 Bidirectional Encoder Representations from Transformers (BERT)-based approaches, both of which take words, n-grams, and/or sentences as input and produce dense embeddings of these texts. These embeddings are analogous to the image embeddings the proposed infrastructure can store to support pre-processed image analysis. Consequently, if a computational social science scholar has sufficient experience with modern neural text analysis pipelines, general approaches from that space transfer well to these image embeddings. In modern search engines, like ElasticSearch, searching documents using these embeddings is often referred to as "semantic search." Alternatively, if a scholar is only familiar with more traditional text-processing techniques, methods exist to align standard bag-of-word methods to image analysis via object recognition and image captioning. Again, the proposed infrastructure could centralize this processing by storing both dense image embeddings and text-based representations of an image via its captions.