Cumulative Cues: Identifying Journalists on Twitter

1st Author Name
Affiliation
City, Country
e-mail address

2nd Author Name
Affiliation
City, Country
e-mail address

3rd Author Name
Affiliation
City, Country
e-mail address

ABSTRACT

Much research has been devoted to tracking the changes to news production and distribution since the widespread adoption of the Internet and social media platforms. An important step to shedding light on these relationships is the ability to distinguish journalists from non-journalists. This task is not straightforward. Relying foremost on data within user profiles, we manually classified 20,662 Twitter accounts that tweeted about a single newsworthy event. Here, we describe our coding process—how it evolved to address specific challenges in the data, and what those challenges and adaptations suggest about the nature of the task, both for researchers and everyday users. Reflecting the messy nature of contemporary digital journalism, we introduced categories for coder uncertainty and content ambiguity. Provisional rules defining journalism were continually challenged by real exemplars. This speaks to the challenge of manual detection of journalists by Twitter users as well as that of systematic detection of journalists at scale. However, those operating journalism accounts can make it easier to identify their role by adopting certain communication strategies.

Author Keywords

Social computing; news production; qualitative mixed-methods; community of practice; Social media; journalism; Twitter.

ACM Classification Keywords

H.5.3 Information Interfaces & Presentation (e.g. HCI): Groups & Organization Interfaces: Collaborative computing, Computer-supported cooperative work;

INTRODUCTION

Historically, journalists have played a distinct role in producing and disseminating information about newsworthy events. This role is undergoing substantial evolution in conjunction with the adoption of social computing tools like Twitter, which ostensibly put journalists and non-journalists on equal footing as potential information producers and sharers. To empirically understand what (if anything) distinguishes the work of journalists in online spaces, we must first identify journalists on Twitter. This important step is less straight-forward than it appears.

Focusing on a single newsworthy event, we sought to distinguish journalists from non-journalists, using visible signals of membership in a Community of Practice of Journalism (CoPJ). Reflecting the complexities of news work in social computing, 18% of the accounts that we ultimately determined could be associated with CoPJ were marked during our process as difficult to distinguish (compared to 4% of non-journalists).

In this paper we: 1) Articulate a method (based on heuristics) for manually distinguishing between journalists and non-journalists on Twitter. 2) Reveal how difficult it is to make these distinctions, providing insight into the nature of information mediation online. 3) Note a few strategies we observed that aided identification. Though not applicable in all cases, these techniques could be more widely adopted, mitigating some sources of confusion we encountered.

FINDING JOURNALISTS ON TWITTER

Many studies that look at journalists' work on social media start with a pre-identified set of 'known' journalists. A limitation of this approach is that it limits our ability to systematically look at journalistic activity of all actors. That is, a case study of a particular journalism organization's work on Twitter related to a newsworthy may be quite informative, but it may not tell us much about *overall* journalistic activity on Twitter for that event.

Distinguishing between journalists and non-journalists across a corpus of tweets about an event affords analysis of the role of journalism relative to that event. A few studies have attempted to systematically categorize journalists on Twitter. Bagdouri and Oardand [2] and Bagdouri [1] use "seed" sets of pre-identified journalists combined with

Paste the appropriate copyright/license statement here. ACM now supports three different publication options:

ACM copyright: ACM holds the copyright on the work. This is the historical approach.

License: The author(s) retain copyright, but ACM receives an exclusive publication license.

Open Access: The author(s) wish to pay for the work to be open access. The additional fee must be paid to ACM.

This text field is large enough to hold the appropriate release statement assuming it is single-spaced in Times New Roman 8-point font. Please do not change or modify the size of this text box.

Each submission will be assigned a DOI string to be included here.

journalism keywords to identify additional journalists who share common characteristics with the initial set of journalists. This approach uses social network relationships and mentions of pre-identified journalists to identify potential journalists. Linguistic similarities including journalism keywords are then used to categorize journalists among the potential candidates. This approach has merit, for example, it may be helpful for identifying candidates for "white lists" of potentially credible news sources. However, this method may not help an average Twitter user to distinguish a journalism account from among the crowd. It also may have limited applicability, as in our case, where we wish to characterize the work of many kinds of journalism across a large social media data set. To use this approach we would have to start with a pre-identified set of journalists to which we could compare.

De Choudhury et al. [3] also systematically classify journalists on Twitter. They make an important distinction between organizations and individuals. They further refine the individual category by distinguishing between "journalists/media bloggers" and "ordinary individuals." These are helpful and meaningful distinctions, but we would like to distinguish *between* journalists and media bloggers as these monikers suggest different kinds of people may be performing the work traditionally taken on by journalists. Additionally, since this work was published (2012), there has been an increasing rise in the prominence of news of questionable provenance as aspirational journalists, click-bait profiteers, and even disinformation actors become increasingly difficult to distinguish from traditional journalism. Thus, we felt it important to revisit the issues of how journalists are classified. Importantly, none of the papers that take on journalism classification squarely address the issue of what journalism is—an issue that we found ourselves recursively revisiting in our qualitative coding process.

METHODS

DATA

This research focuses on Twitter accounts that participated in information-sharing around a single newsworthy event: the 2014 Oso Landslide. This tragic crisis event took 43 lives and destroyed a rural neighborhood in Washington state on March 22, 2014. As the largest mass-fatality slide in U.S history, the Oso Landslide garnered attention from large and small news organizations across the globe. Conversely, compared to other disasters that gain international attention, the slide was small in scope, damaging only a square mile and generating a relatively small digital footprint. For this reason, this data set in particular lends itself to a comprehensive exploration of Twitter activity without the need for sampling, thus enabling us to map a broad range of journalistic activity respective to a single event.

We purchased a collection of tweets containing event-related keywords, hashtags, and locations posted from one day prior to the slide to three weeks after. From this collection of 986,826 tweets, we scoped to tweets that used one of two event-specific hashtags (#OsoStrong or #530slide) determined through prior research to be relevant. This dataset consists of 78,409 tweets and 20,662 accounts.

CODING PROCEDURES

Next, we attempted to categorize each of these accounts as being journalists or not. Anticipating difficulties in making categorical determinations, we chose a coding process that enabled us to identify and closely examine accounts that were categorically problematic.

All 20,662 accounts were independently categorized by two coders as either "journalist", "not a journalist" or "unsure." To make an initial determination, each coder reviewed information from the user profile as it appeared at the time of the event: user name, description, number of posts, number of followers, number following, URL, and geographic location.

When a disagreement arose between coders or both coders marked an account as unsure, additional steps were taken to make a determination. These included, as needed, a review of current Twitter account activity, LinkedIn profiles, Wikipedia articles, "About Us" and "Contact Us" pages on websites, Twitter ID lookups, et al.

For the first third of the set, all four coders reconciled in person to consensus (5749 accounts, 422 disagreements, 16 unsures). At this point, we saturated on rationales for making a determination and sources of confusion. For the remaining ⅔ of the set, we used a third coder to arbitrate coding disagreements. "Unsures" were assigned to two coders who investigated the account and converged on a determination. "Unsures" were ultimately re-categorized as "yes", "no", or "ambiguous". "Unsure" indicated uncertainty on the part of coders after reviewing only profile information. Ambiguous indicated that even with additional web searches and discussion, a determination was inconclusive.

PROVISIONAL RULES FOR IDENTIFYING JOURNALISTS

Through a grounded approach, we started with provisional rules for categorizing journalists based on previous literature and consultation with a domain expert, a former newspaper reporter. We then modified our categorization heuristics based on issues we encountered during coding.

An association with a journalistic community of practice

An overarching criterion guiding all our decisions was this: *Does this account associate itself with a Community of Practice of Journalism (CoPJ)?* Following the concept of Community of Practice as outlined by Wenger [7]—membership is learned and performed through interaction with other members—we identified several ways that an account might signal such membership. We include the following accounts: anyone who makes an identity claim of currently being a journalist; anyone learning membership by training led by journalism educators (e.g., a claim of studying journalism at a university) or leading such a training (e.g., claiming to be a journalism professor); anyone claiming to be learning on the job (intern at a news organization). Membership in the CoPJ carries credibility and accountability, as training involves learning to follow a code of ethics (e.g., Society of Professional Journalists Code of Ethics), which acts as a safeguard against bad information.

Because the concept of Community of Practice emphasizes the internalization of ways of thinking and particular practices, we considered it appropriate to also include accounts that made claims to prior work as professional journalists (e. g "recovering journalist" or "Former News Reader for Sky News"). For similar reasons we also include professional organizations that are made up of journalists or service them (e.g. @traumajournos, @poynter).

Does this account participate in journalistic activities

Our second guiding question was: *Does this account have to do with news or participate in journalistic activities?* We excluded those who stated they were associated news organizations but had jobs titles such as marketing or IT that were not directly involved in newsgathering or news production. Journalistic activities can be considered in a hierarchy [4]. Witnessing a newsworthy event and sharing information about it is at the bottom of the ladder, followed by other distinct types of activities, including seeking corroborating evidence, interviewing people, vetting sources, confirming information before sharing it, analyzing what happened, providing context.

Is this journalism?

Though several of the studies we reviewed give no definition of journalism, as we struggled to reach consensus as a team over difficult to categorize accounts, we found it necessary to have a definition to work with. A journalist is someone employed to regularly engage in gathering, processing, and disseminating information to serve the public interest [5]. Online publishing platforms and changes in the industry make any definition problematic, however. That's why our guiding questions for the detection criteria focused on defining what activities constitute journalism.

Journalism is most distinct from other forms of mediation when it can be seen as playing a watchdog role in the public interest. For example, a trade publication that covers legislative issues might be journalism, while one that does not is harder to categorize as such. Journalists cultivate sources and audiences, mediating between domain experts and the public. Yet, many meteorologists (a kind of domain expert) work in news. In short, distinct criteria for what constitutes journalism quickly collide when confronted with real-world examples. We found two criteria to be most helpful: 1) When reviewing possible journalism sites, we considered whether some content on the site could be seen as playing a watchdog role. 2) We also sought out evidence of some original work in reporting, curation, or editorial framing. This very basic criterion cut out numerous aggregating accounts (often bots) that purport to be news.

Additional Cues

We also considered all data points in the user profile including follower/following ratio, number of posts, length of time the account has been active, whether the account was verified, images, URLs, account location, and so forth.

FINDINGS

Coders were most confident in determining that an account was associated with journalism when multiple signals of membership in the CoPJ were present, language used to signal membership was unambiguous, and a cross-platform presence was apparent. However, many Twitter accounts that we ultimately determined to belong to journalists signaled membership in a more muted way, leading to uncertainty and disagreements in how an account should be categorized. We first offer an example of an account that was easy to identify, then explain some of the issues that led to uncertainty or disagreement in determining an account is associated with CoPJ.

Cumulative Cues Lead to Confident Identification

Of the 2244 accounts we identified as journalists, 1845 (82%) fall into the category of easily identifiable. In these cases both coders independently identified the account as a journalist without expressing uncertainty about their decision in the first round of coding. One exemplar of an account coders confidently identifiabled as journalist is @AndreaWoo. Woo makes an unambiguous claim to membership in the CoPJ by using the word "journalist" in her account description. This account has been active a reasonable length of time (since 2009), which is not conclusive, but adds credibility. Number of tweets to date (28.4k) and number of followers (17k) suggest that this account is quite active and plays a role in informing others on Twitter. The number of accounts that Woo is herself following is substantially lower than those following her (1877). Such a follower/following ratio has been suggested by other researchers as appropriate for a journalist.

Importantly, Woo associates herself with an outlet that sounds like a news organization, including both an email address and an associated URL, theglobeandmail.com, in her account description. Clicking on the URL takes us to a page on the Globe and Mail's site that loads all of Woo's stories for the Globe and Mail in reverse chronological order. Though her title changes on the website to "News Reporter", identity claims presented on each platform are coherent and consistent. But, we can also go beyond claims made to review Woo's work if desired. In one click we have moved from Woo's Twitter account description to a substantive corpus of her work. An additional click takes us to a story written by Woo, complete with her byline and date it was published.. Many other attributes of the Globe and Mail website signal that it is a legitimate journalism outlet. For example, an easy-to-find functional staff directory, listed physical address and phone number.

Ambiguous Cues

Ambiguous Job Titles

We encountered very few cases where we contested the identity claim of an account that described themself as a "journalist." However, numerous news-related job titles are ambiguous. Many broadcast jobs, such as "host", can only be inferred as journalism-related if one can determine the production they are associated with has a news component. For example a host of NPR's Morning Edition is a journalist. The host of a music show is probably not. Some journalism accounts used no journalism keywords to describe themselves, instead simply affiliating themselves with an outlet. For example, the account @JamesQ13Fox gives no user description. It's up to the reader in these cases to infer if the outlet is journalism and what role the account owner plays in the outlet. In other cases, journalists used insider language such as "copy guy" or "science writer" that was not recognizable to one or more members of our team as a journalism job title. Additionally, many journalism job titles sound similar to those in related professions. We continually ran into confusion for job titles that overlapped with public relations, marketing, government relations, and entertainment. This suggests that journalists could amplify their association with CoPJ for non-journalists by assuring one or more widely recognized journalism keywords appear in their profile.

Is This is a News Organization?

We were most confident that an account was associated with journalism when it directly affiliated itself with a news organization. However, many journalists who have an affiliation choose not to directly name the affiliated organization in their Twitter profile or link to it directly. We acknowledge, journalists may have good reasons for doing so, Yet, the more immediately and directly we could link an account to what we could determine was a news organization, the more confident we were in making a quick determination. An inescapable fact of this process is its reflexive nature. The more social or cultural distance there was between a coder and an organization, the more difficulty we had making this determination. In our data set, sports fans, sports journalists, sports bloggers and sports podcasters are all well represented— all tweeting about a landslide. No one on the coding team follows sports. Therefore, we found it difficult to tease apart the positionality of different sports actors. Certainly, some fans are journalists, but which ones? Certainly, some sports sites are journalism, but how to tell? In relation to tweeting about a landslide, do those distinctions matter? In general, the smaller the scope of an outlet (be it geographic or niche audience) or the more specific the slice of the public it targets (realtors, engineers, insurance agents), the more difficulty we had determining if the outlet was doing journalism.

Yet another confounder is when individuals who claim to be journalists work for organizations that do not claim to be news organizations. For example, are science writers who work for universities considered journalists? They serve the public interest by sharing important scientific knowledge on behalf of the university, which suggest they are journalists. They have been trained in the CoPJ and are engaging in journalistic activities. But can they truly play a watchdog role when their job is to cast the university in the best light —a PR role? For example, if the university becomes embroiled in a scandal, the public interest and the interest of the university will diverge. This example demonstrates the limitations of any comprehensive attempt to define journalism and journalists.

Cross-Platform Cues

Perhaps most surprising among coders (3/4 of whom are digital natives) was the weight given to a cross-platform presence. We had a harder time categorizing digital-only outlets as journalism. Some blogs perform journalism and some are written by people who identify themselves as journalists. A digital-only presence is more attainable to those wishing to play a role in news production and dissemination: the aspirational news producer with no affiliation to CoPJ, genuine community and citizen journalists who educate and inform the public albeit with little direct connection to CoPJ, marketers and disinformation actors are all present. Therefore mixed signals abound. Even highly recognizable outlets such as @Mashable and @BuzzFeedNews confounded us by avoiding describing themselves with language that would clearly link them to the CoPJ. Yet, in some cases individuals working for those outlets do describe themselves as such. Therefore, for digital-only outlets, we found ourselves relying in tandem on whether we could both view news content on the site and we could readily identify humans as publishers and content producers—a standard of traditional journalism. For example, could we find contact information such as a physical address? Did those named as contributors have a digital presence beyond social media sites? Could we tie them to bylines of actual news content or to other roles indicating that they participate in journalistic activities? In some cases, we investigated biographical details of those individuals, looking for a presence outside of the sites we were unfamiliar with. Occasionally, even then it was hard to make a determination. Archives on many news sites, such as small-market news organizations, are notoriously unreliable. We observed many cases where signals of journalism work would be by strengthened by improving cross-platform cues...

Finally, although newspapers are canonical news organizations, coders sometimes had difficulty identifying traditional (or "classic") newspaper names, thereby dismissing them as not journalism. On occasion, this was compounded by journalists and outlets abbreviating the publication names. In addition, names of legacy outlets are often mimicked by aspirational news producers, aggregators, marketers, and disinformation actors. Thus, if a legitimate legacy outlet was not already known to a coder, it raised uncertainty for the coder. Occasionally we came across an outlet that anticipated our confusion by giving the kind of details that signal their position within a community. E.g., @ktivnews describes itself as "*Serving Siouxlanders since 1954, we pride ourselves on being Siouxland's NewsChannel!"* When descriptions include metrics (in this case years of operation) and an explicit statement on the public served, we found it reduced our confusion because it helped distinguished these journalists from aspirational actors and news aggregators.

Neutral Cues

A number of cues identified by previous research as associated with the CoPJ identified were only marginally helpful when looking at accounts on an individual basis. Kamps [6] claims 25% of verified accounts are journalists (though journalists are aggregated with "media" in that study). However, many of the journalism accounts reviewed were not verified. Verified status only swayed our decision in a few cases.

Likewise, the ratio of followers to following has been used in machine learning studies to detect journalists. However, 931 of the 2,244 journalists (41%) had an inversion of the ratio, where their friends count was greater than their following. This suggests that the ratio may better detect accounts *performing* as news sources within the Twittersphere—some significant portion of which may be journalists, rather than a method of detecting journalists.

DISCUSSION & CONCLUSION

We were most confident designating an account as a journalist when multiple signals of membership in the CoPJ were present, language used to signal membership was unambiguous, and a cross-platform presence or journalistic activity was readily apparent.

Twitter is widely recognized by journalists as an important tool for gathering and disseminating news. That means that Twitter users can rely on the platform to find credible information quickly. However, the proliferation of misinformation and disinformation actors on social media makes it challenging for Twitter users to detect trustworthy information sources, especially when it comes to distinguishing legitimate journalism from what has become known as fake news. We reviewed 20,662 Twitter accounts that tweeted a single newsworthy event and attempted to identify accounts associated with journalism. The ambiguities and challenges we encountered in the process revealed a need to define what journalism is and who can be considered a journalist. Our first contribution is methodological: We outline in this paper how we arrived at our definition and tested accounts against it. Our second contribution is an implication for design: We identified that certain cues within a Twitter profile work jointly to signal whether the account belongs to a journalist. These cues suggest certain strategies journalists can use to ensure that Twitter users can identify them quickly and easily.

REFERENCES

1. Mossaab Bagdouri. 2016. Journalists and Twitter: A Multidimensional Quantitative Description of Usage Patterns. In *ICWSM*, pp. 22-31.
2. Mossaab Bagdouri, and Douglas W. Oard. 2015. Profession-based person search in microblogs: Using seed sets to find journalists. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Managemen*t, pp. 593-602.
3. Munmun De Choudhury, Nicholas Diakopoulos, and Mor Naaman. 2012. Unfolding the event landscape on twitter: classification and exploration of user categories. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pp. 241-244.
4. Steven Myers. 2011. Why the man who tweeted Osama bin Laden raid is a citizen journalist. *Poytner.org.*
5. Jonathan Peters and Edson C. Tandoc Jr. 2013. 'People who aren't really reporters at all, who have no professional qualifications': Defining a journalist and deciding who may claim the privileges. In *Proceedings of NYUJ Legis. & Pub. Pol'y Quorum 2013* pp. 34-34.
6. Haje Jan Kamps. 2015. Who are Twitter's verified users? Retrieved July 27, 2017 from https://medium.com/@Haje/who-are-twitter-s-verified-users-af976fc1b032
7. Etienne Wenger. 1998. *Communities of practice: Learning, meaning, and identity*. Cambridge University Press.