Journalistic Source Discovery: Supporting The Identification of News Sources in User Generated Content

Yixue Wang
yixue.wang@u.northwestern.edu
Northwestern University

Nicholas Diakopoulos
nad@northwestern.edu
Northwestern University

ABSTRACT

Many journalists and newsrooms now incorporate audience contributions in their sourcing practices by leveraging user-generated content (UGC). However, their sourcing needs and practices as they seek information from UGCs are still not deeply understood by researchers or well-supported in tools. This paper first reports the results of a qualitative interview study with nine professional journalists about their UGC sourcing practices, detailing what journalists typically look for in UGCs and elaborating on two UGC sourcing approaches: *deep reporting* and *wide reporting*. These findings then inform a human-centered design approach to prototype a UGC sourcing tool for journalists, which enables journalists to interactively filter and rank UGCs based on users' example content. We evaluate the prototype with nine professional journalists who source UGCs in their daily routines to understand how UGC sourcing practices are enabled and transformed, while also uncovering opportunities for future research and design to support journalistic sourcing practices and sensemaking processes.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**.

KEYWORDS

Computational journalism, crowd-sourcing journalism, citizen journalism, user-generated content, journalistic sourcing, sensemaking

ACM Reference Format:
Yixue Wang and Nicholas Diakopoulos. 2021. Journalistic Source Discovery: Supporting The Identification of News Sources in User Generated Content. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3411764.3445266

1 INTRODUCTION

Journalistic sourcing, or seeking timely information for reporting, is a key element to successful news production. Nowadays, user-generated content (UGC), including everything from forum posts, to social media, comments, weblogs, and callout surveys, serves as a pool of initial sources and tips for a new or ongoing story, and can help journalists gain new contacts or receive information for follow-up stories [78]. For instance, comments, as a form of UGC, offer not only a forum for critique but also a diverse set of alternative opinions, perspectives, and interpretations from a range of community sources, which can lead to stronger and more rigorous journalism [24]. Callout surveys (i.e., asking audiences to respond to a specific request[1]) are another form of UGC that engages in targeted outreach via multiple channels, including emails, SMS, website or Google forms and can enable voting, witnessing, sharing personal experiences, tapping specialized expertise, completing a task, and more generally just engaging audiences [47]. More broadly, social media, where large volumes of UGC diffuse, enables journalists to interactively contact readers and promises to enable more direct connections to elites, the voices of the people, and regions that are difficult to access [74]. To unify these various types of UGCs, we use the broader conception and typology of "audience material" [77] in this study.

The deluge of information brought by UGCs has provided opportunities for newsrooms, but also unanticipated challenges. Prior HCI research suggests multiple challenges with sourcing from UGCs, including determining the credibility, quality, and relevance of UGC while also preventing the spread of misinformation [11, 40]. For instance, one study found that 32 major U.S. news outlets had referred to at least one Russian IRA tweet as a source when reporting partisan opinions from 2015 to 2017 [35]. Respondents in another study suggested that UGCs could undermine journalistic values unless carefully monitored, a practice that doesn't always fit easily within traditional newsroom routines [64]. Current journalistic sourcing practices therefore need to be explored to better support those practices in sociotechnical systems.

Journalists working online typically have less time for researching stories [68], and the online news format puts additional pressure on rapid updates, often reflecting a "relatively loose culture of corroboration" [70]. Online news platforms are in constant competition with other news outlets, sometimes by "fractions of seconds" [69]. Some tools have explored different design approaches and functions to support monitoring and identifying newsworthy information, curating news, and validating news sources from UGC platforms under this pressure [17, 18, 67, 80]. However, there is more limited research on how to design tools that help explore and deeply investigate a range of UGCs in a way that effectively supports varied journalistic sourcing needs and practices. In this work, we focus on text-based UGCs reflecting "audience experiences" (case studies contributed in response to a news story), "audience stories" (story tip-offs from the audience which are not on the news agenda), and

[1]For example: https://www.propublica.org/article/she-was-sued-over-rent-she-didnt-owe-it-took-seven-court-dates-to-prove-she-was-right

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445266

"audience comments" (opinions shared in response to a call to action) [77], rather than on other UGC data types (e.g., user geography, eyewitness images/videos).

To help journalists find interesting leads and newsworthy text-based audience materials, and to introduce UGCs into journalists' reporting process more efficiently via tool design, in this paper we seek to understand UGC sourcing practices, to design tools to effectively support these practices, and to propose broad design suggestions that connect news reporting to information sensemaking processes [53]. We first report the results of an initial qualitative interview study with nine professional journalists who have experience in UGC monitoring and sourcing working at a range of local, national, and international news organizations. Based on the insights from this study, we develop several reporting scenarios as well as design goals to guide the development of a UGC sourcing tool. We then detail the development of this prototype tool and present an evaluation of the system with nine professional journalists who have experience sourcing stories from UGCs. Findings from the evaluation indicate that the prototype is able to help identify desired content for journalists, and indicate several broad suggestions and opportunities for future work in designing tools to support journalistic sourcing and sensemaking processes.

This work offers four main contributions, including: (1) The articulation of journalistic tasks, approaches, and needs related to UGC sourcing practices based on an initial qualitative interview study with practitioners; (2) The development of design goals and a journalistic sourcing tool to enhance journalists' efficiency of sourcing from UGCs using a scenario-based design approach informed by the initial interview study; (3) An evaluation of the sourcing tool with professional journalists in order to examine the utility of the prototype and explore its potential to effectively support journalistic practices; and (4) The identification of design and research opportunities for future tools to support journalists by relating our evaluation results to an information sensemaking model [53]. These contributions help to elaborate domain-specific needs and ways forward to better support journalistic sourcing practices through more advanced computational tools and interfaces.

2 RELATED WORK

In this section, we motivate our research questions by drawing on literature related to how traditional sourcing practices are evolving with respect to the modern media ecosystem, and what kind of current tools support journalistic practices around UGCs.

2.1 The Evolution of Journalistic Sourcing Practices

Journalists often rely on expert knowledge when they gather and verify news or explain and contextualize events [39]. Traditional journalism practices reflect a hierarchical interpretation where only a certain set of professionals and elite stakeholders decide what information to publish to the public [75]. Theories such as gatekeeping (i.e., the process of culling and crafting information into a limited number of messages [62]) and the "two-step flow" model (i.e., messages flow from mass media to opinion leaders to the general public) [30, 62], highlight the hierarchies structured around these gatekeepers—journalists and opinion leaders—and have historically tended to disregard how ordinary people interact with the communication process. The Hierarchy of Influences model suggests reasons for why journalists prefer elite sourcing, outlining how personal relationships, relevance, accessibility, willingness to talk, and credibility are several key elements behind journalistic source selection [20, 33, 61].

But traditional journalistic sourcing practices are also evolving alongside changes in the digital media ecosystem [28, 66]. News production processes have become increasingly intertwined with influences from a broader array of stakeholders, including distribution platforms, algorithms, and audiences [75]. Traditional gatekeeping theory has evolved to include more stakeholders, including UGC platforms and the algorithms behind them [76]. Platforms can work as part of broader sociotechnical gatekeeping practices since they can contribute both to news distribution and news production, acting to influence both consumer attention as well as various gatekeepers prior to publication [16, 26, 51]. HCI researchers have acknowledged these changes in stakeholders and provided journalists with digital creativity support tools to discover and examine creative angles [37] and studied journalists' roles [11] within new and emerging digital news ecosystems. The one-way broadcast era is ending, and the role of journalists is fundamentally shifting away from being the only providers of information to instead being responsible for finding "*credible, authoritative voices in a noisy world*" [10], since unmediated content is not subject to the traditional quality criteria of news publication processes [25] and journalists may collaborate with and coach the crowd in reporting processes on UGC platforms [11]. In this new media ecology, ordinary citizens (or "*real people* " as described by [21, 65]) "*play an active role in collecting, reporting, analyzing, and disseminating news and information*" [4].

The shifts in journalistic sourcing practices within the rapidly changing digital media ecosystem have created new opportunities for newsrooms to tap into UGC in news production [1]. For instance, mobile phones have created easy ways, such as SMS messages and customized mobile applications, for newsrooms to solicit UGC submissions (specifically video and photo submissions) from their audience [72]. HCI Research in mining UGC (primarily focused on social media due to their low barriers and large volume) has shown different design tasks in the news cycle, from news discovery, to curation of news, validation of content, and newsgathering dashboards [80]. Journalists may listen to audiences for a number of reasons, such as *witnessing*, *observing personal experiences and stories*, and *tapping into specialized expertise* [5]. Audiences can share what they see during a breaking event, natural catastrophe, or through life experiences, and contribute unique knowledge. For instance, the BBC has used UGCs as a form of tip-off in order to add depth, weight, and detail to a story, as well as to write investigative articles, which come from people who "*have the first-hand experience of the story or at the location where the story is taking place*" [78]. Journalists may include information from social media because it's newsworthy or as a way to support or exemplify a story. In some cases, individual tweets or interaction between various agents on Twitter can trigger news coverage [6]. UGCs from social media have been used to add flavor to an anecdote from someone directly involved [5].

Researchers foresaw that journalistic sourcing cultures would become more open to diverse and alternative sources due to the abundance of UGCs [27, 54]. And many newsrooms have already integrated UGC content into news reporting practices, including *ProPublica*, *The Washington Post*, *The New York Times*, and others [41]. Over the past few years UGC content has been increasingly accepted as a legitimate source of information in mainstream U.S. media outlets [41, 57]. One study found that from 2013 to 2017, 9.2% of news articles cited at least one social media post as a source [57]. Another comparative study of mainstream media outlets also found an increase in social media sources among *The New York Times*, *The Guardian*, and *Süddeutsche Zeitung* in recent years [74]. However, the verification of online sources is full of challenges [33], including constantly changing environments, lack of action and cooperation with data owners, and lack of reliable reach data [40]. Research has also found that some types of news, including breaking news and crisis news, benefit from online sources to provide faster news updates and include more diversity of voices shortly after an event [11, 66], while the usage of UGCs in non-breaking news events is still under-explored. An in-depth understanding of current sourcing practices across newsrooms, including the various practices for leveraging different types of UGCs (e.g. comments, social media posts, or survey call out responses) as well as the ways in which UGCs fulfill various sourcing needs, is still largely missing. This leads to our first research question:

- **RQ1**: *What are journalists' sourcing needs and practices as they seek information from UGCs?*

2.2 Tools to Support Journalistic Practices Around UGC

A range of tools and interfaces have been developed to assist journalists with gathering information and identifying newsworthy items from UGC, from event and eyewitness detection to source curation and verification.

Many tools aggregate and visualize UGC content to help users interactively identify newsworthy information. *Vox Civitas* is a tool to help journalists and media professionals identify news from aggregations of social media content around broadcast events. The tool presents tweets aligned to a video clip, a topic timeline, a message volume graph, a sentiment timeline, and salient keywords related to the broadcast event over time [18]. *RevEx* is another example of an analytic platform built to help journalists search and visualize crowd-sourced reviews on healthcare providers from Yelp using keyword and faceted search [19]. Tools such as *SocialSensor* as well as commercial apps including *Geofeedia* and *Spike*, have been found to be helpful for journalists in terms of finding original news, detecting trends, and verifying information from social media platforms [67]. *Seriously Rapid Source Review (SRSR)* incorporates advanced aggregations, analytic computations, and cues that can be helpful for journalists to find interesting and trustworthy information sources, including eyewitnesses, who have potential to provide relevant content [17].

Other tools have been designed to help journalists with real-time or geographically-oriented monitoring. *CityBeat* is a system designed to find possible news events by analyzing real-time geotagged information from social media and visualizing the stories for journalists. A deployment of the system into professional newsrooms found that it favored certain types of news events and was biased to certain populations and interests, highlighting a key challenge in identifying representative content from UGCs [58], a challenge that emerges again in the current work and which we address in our prototype design. Systems can also help journalists monitor UGCs for new and emerging events by automatically detecting surprises, anomalies, and changes [15] or by using a predefined set of keywords or hashtags to collect data [80]. *Hotstream* is an early approach to group and detect breaking news from Twitter, ranked based on popularity, reliability factors, and keyword (e.g., "#breakingnews") search [52]. Another system, *Tracer*, is an automated Twitter monitoring system deployed at Reuters News, which helps journalists both detect and evaluate breaking news events [34].

Oftentimes tools focused on social media emphasize validating sources and fighting against rumors and fake news. For instance, the *Pheme* project uses machine-learning techniques to help journalists observe the veracity of potential rumors propagating on social media, find reliable sources, and recognize how rumors surface and develop [12]. *RumorLens* provides a dashboard tool for journalists to visualize data from Twitter, identify rumors as they appear, and assess their speed of propagation and the extent to which these rumor messages have been subject to correction on the platform [56]. *TwitterTrails* [42] can identify when rumors first appear on Twitter, how they propagate, and the ways they might have been corrected. *The Dashboard* [71], designed as a social media verification tool, can capture temporal fluctuations and surface stories from Twitter data for journalists.

Many aspects of supporting journalistic practices around UGC sourcing are explored by researchers and journalists as discussed above. There are clearly a wide range of tasks and journalistic contexts in which UGC is relevant. Perhaps the closest to the current work is the *CommentIQ* system [49], which demonstrated the journalistic need to be able to identify sources from comments. Yet, that work was still narrowly focused on comments rather than the wider array of UGCs which we consider here, and it was tailored to comment moderation tasks rather than more broadly considering how reporters and editors may want to find sources from UGC. In contrast to many of the other tools cited above, the current work focuses on different use scenarios that are less event-oriented or geared towards breaking news, and more driven by other journalistic contexts and goals in making sense of UGC content from sources such as comments or survey call-outs. In this work, we adopt a wider lens and build on the insights of many of these prior tools to inform the design requirements of our own prototype, which we elaborate further in Section 4.2. Through the design and evaluation of this tool with moderators, reporters and editors, we strive to further develop human-centered insights that can inform the design of effective journalistic sourcing tools by contextualizing our evaluation results in terms of information sensemaking (see Section 6.2). This leads to our second research question:

- **RQ2**: *How can we design tools to effectively support journalistic sourcing practices from UGCs?*

3 STUDY OF UGC SOURCING PRACTICES

In this section, we describe our effort to address RQ1 by undertaking a qualitative interview study with nine professional journalists. We first describe the procedure and methods for the study, and then elaborate our findings on UGC reporting approaches and broader journalistic information needs met by UGCs. These findings provide insights that help inform the design goals for our prototype described in Section 4.

3.1 Study Methods

We collected data from early 2019 to early 2020 in the form of nine in-depth interviews with full-time professional staff journalists working within U.S. newsrooms in editorial roles that incorporated UGC in their practices. All interviews were conducted via voice calls using a semi-structured interview guide. The interview guide consisted of a set of 10 initial questions, with additional follow-up questions if relevant. The initial questions covered internal workflow around UGC sourcing practices, the motivations behind using UGCs as news sources, examples of participants' UGC sourcing projects, and challenges and problems with journalistic sourcing from UGCs.

We first identified potential interviewees through contacts from professional journalists known to us for their activities in community engagement. We also searched Google, LinkedIn, and Twitter with relevant job titles (e.g. "community editor" and "audience editor") in news organizations. Finally, we used snowball sampling by asking interviewees for references to other relevant practitioners. These recruiting methods resulted in nine interviewees (I1 ... I9), of which three held reporter roles, six held editor roles, and two held moderator roles (two interviewees had both editor and moderator roles). Six interviewees identified as female, and three as male. Interviewees were from regional topical newsrooms (N = 2), local public radio (N = 1), and some of the biggest (inter-)national news organizations in the U.S. (N = 6). Interviews ranged from 36 to 60 minutes (Median = 52 minutes). We include Table 1 for more information about each interviewee, including their role and organizational context.

Audio recordings of the interviews were transcribed and then qualitatively analyzed using a thematic analysis approach [22], incorporating iterative coding, constant comparison, and memoing [9]. Transcripts were first open coded to identify salient excerpts and organize those into preliminary categories. As we coded, excerpts were iteratively grouped according to the emerging categories while constantly comparing them to ensure they formed coherent thematic ideas. Memos and diagrams of the themes were written to help make sense of and reflect our understandings of UGC sourcing and, because analysis was ongoing during the interview process, to help conceptually shape future interviews. Throughout the process, analyst triangulation was practiced by discussing the themes and memos between the co-authors.

3.2 Findings

The outcome of our analytic process resulted in the identification of two high-level approaches or practices related to how journalists use UGC, as well as a number of specific information needs. In the following subsections we first present the two high-level sourcing approaches identified, *deep reporting* and *wide reporting*. We then elaborate what journalists look for in UGC content based on the main themes that emerged from our interviews, specifically we introduce journalistic information needs for personal experience and expertise, community responses and trends, questions, and different viewpoints and opinions. These findings motivate our design to support journalistic UGC sourcing discussed in Section 4.

3.2.1 Two UGC Sourcing Approaches. Even though different newsrooms have different approaches when using UGCs in their reporting because of different editorial values and needs, we observed two distinct approaches towards sourcing that emerged from the interviews: *deep reporting* and *wide reporting*. These approaches are not mutually exclusive and can be complementary when journalists source UGCs. For both deep and wide reporting approaches, it is essential for journalists to have users' contact information to validate information with the sources. As stated by I5, journalists "*have to verify, and obviously sometimes online sources can make it more of a challenge to verify,*", so they try to acquire respondents' emails and/or phone numbers by either asking them directly on whatever platform they contribute to, or asking them to fill out such information in their callouts and surveys. We characterize and contrast the two approaches in Table 2 and detail them more in the following subsections.

Deep Reporting. Deep reporting involves finding and investigating specific people of interest with a particular perspective or expertise that journalists want to pursue. It tends to be more hypothesis-driven. To write a feature story or an extensive investigation, journalists must build strong connections with their sources found from UGCs. Whoever contributes the desired UGC posts will be invited to "*join the investigation and join reporting.*" (I1) Journalists then interview interesting contributors, listen to them to shape the investigation and may eventually use them as sources in their final report. I4 described an interesting workflow example using callouts to find their leads:

> We were trying to find people who participated in the study or may have been part [of it]. But we were not getting the records that we needed ... So we put out a callout basically ... looking for people who are affected by this study, and we got an email from a woman ... We ended up working with her really really closely on retelling the story of her son's experience through her journal entries.

I1 shared a different workflow that she uses when monitoring comments. In this type of workflow, journalists are more likely to write a follow-up story instead of starting a new one since they were already reviewing comments from a previous story:

> If someone leaves a comment and says I have experience with this company and I XYZ ... We're not going to, we can't [directly pull out their contact information]. If our journalists want to contact them, then they'll go into the comments and say hey, you know we'd love to talk to you. Here's my email address if you'd like to reach out there.

Table 1: Interviewee Details

Interviewee	Role/Title	Type of Organization
I1	Head of Audience	International News Outlet
I2	Chief of Staff and Associate Director of Policy	U.S. State Topical News
I3	Director of Growth	U.S. State Topical News
I4	Engagement Reporter	Large Non-Profit U.S. News Outlet
I5	Community Editor	Large U.S. News Outlet 1
I6	Director of Training, Professional Development and Operations	Local Public Radio
I7	Editorial Director	Large U.S. News Outlet 2
I8	Journalist and Partner Manager	Large Non-Profit U.S. News Outlet
I9	Community Editor	Large U.S. News Outlet 3

For this approach, the representativeness of the person within the community doesn't necessarily matter that much since journalists want to "go deep" to investigate a personal experience exposed by the UGC post. Finding UGCs is just a first step in a much longer reporting process. The contributors may even be involved in the actual writing process, providing deep insights and helping to shape the story. The content of the UGCs works more as an identifier of someone who then becomes a source, instead of acting as a quick piece of content that might get embedded or quoted.

I6 explained in detail why his newsroom prefers this reporting approach. For his newsroom, the UGC volume they receive from platforms is small enough that his newsroom can review them all, which allows them to be focused in the reporting and build stronger connections with contributors. The newsroom editorial values lead to this preference too: "*[We] want to start from a very different relationship of editors with the public that they're serving* ." His newsroom felt very uncomfortable deciding what they were going to be reporting on all the time without actually checking with members of the public who were supposed to be consuming it. By introducing the community to their investigative process, even going so far as to include them in decisions about what direction to take, the real relationship with the person who contributed "*makes a huge difference in the editorial outcome,*" and the public is able to help journalists "*direct their journalism.*"

Wide Reporting. Wide reporting involves exploring the overall community reactions through collecting, analyzing, and interpreting representative, unanticipated, and dissenting responses. It tends to be more open-ended but is still shaped around journalistic goals or interests. If journalists are writing a wrap-up or a community response article, they need to find the most interesting stories out of many UGCs and make sure they are representative and reflect editorial values by incorporating dissenting opinions to achieve a variety of viewpoints. Stakeholders, locations, and political leanings may be considered. In the end, they try to look for a range of stories, perspectives, or viewpoints on an issue from the whole community. In this case, journalists may use direct quotes from UGC, and they focus more on organizing the identified personal experiences and quotes instead of writing articles closely with the contributors. I3 explained their workflow around UGC sourcing, specifically when they try to report on a specific topic from different angles (i.e., overall questions, missing voices, etc.) by reaching out to their target audiences via callout surveys:

> The first thing we do is say, "What question would we like to ask a group of teachers or a group of parents? What piece of information or what voice is missing from this article?" So then that helps us decide what the actual questions are going to be in the set. And we usually try to keep it pretty brief... Then we actually create the survey and then we go over to our e-mail platform... we start following up with people if they've given us their name or email. And then we just incorporate the responses we're getting both from the poll and from our follow up calls [with callout responders] into the article.

Here is another wide reporting workflow example when monitoring a different UGC format (comments) from I5, where journalists consider as many perspectives as possible:

> We did a piece about young evangelicals, and it got a ton of comments. If we did a reader roundup, we would sit down and ... [count] X amount of readers who voiced this opinion, and this is what Sarah in Oklahoma said. And we would go through and take really good comments and just give a general synopsis of the different things that we've done. We would pick ... positive comments and interesting ones and controversial comments. And this is what the dissenting voices said. And that's one kind of follow up.

For both callouts and comments, this type of article writing workflow requires a much larger volume of UGCs than the previous workflow, and representativeness matters to a much greater extent. Journalists try to find diverse respondents and reflect that diversity in their writing. Not only do journalists care about personal stories, but they also care about dissenting opinions and community trends. The large volume of UGCs here is not a disadvantage. On the contrary, it allows journalists to do analysis based on UGCs, notice community interests more quickly and easily, and report the

Table 2: Characterizing and Contrasting The Two Approaches

	Deep reporting	**Wide Reporting**
Scope of Interest for Sources	Specific people of interest	The broader community
Scope of Interest for Content	More hypothesis-driven	More open-ended, but still shaped around journalistic goals or interests
Story Type	A feature story or an extensive investigation	A wrap-up or a community response article
Data Size	Could be smaller	Typically a larger volume
Content Representativeness and Diversity	Not necessary	Essential
Connections with Contributors	Stronger connections to direct investigation and verify information	Weaker connections for verification of quotes or excerpts

findings in categories found from UGCs. I5 specifically mentioned that "*just because one person writes to complain about something doesn't mean that it's actually impacting a large number of people,*" and thus they really look for "*trends and patterns*" in the content.

3.2.2 What Do Journalists Look for in UGC Content? Prior work has shown that journalists may listen to audiences as part of their crowdsourcing practices for reasons such as witnessing, sharing personal experiences and stories, and tapping into specialized expertise [6]. In our interviews, we find that journalists engage with their audience across the full news production cycle, not only for crowdsourcing audiences' experiences and expertise, but also to meet other informational and journalistic needs. Interviewees emphasized the importance of personal experience and expertise, but also shared that they use UGCs to find new stories via tips from posts and to shape follow-up stories with audience responses and trends, questions, and different viewpoints. While journalists expressed different forces on their use of UGCs, including the drive for more traffic and engagement, use was most often framed in terms of the journalistic purposes of supporting and interacting with the community as a public service. Note that these information needs can be sought in both approaches and are broadly applicable in UGC sourcing practices.

Personal experience and expertise. When journalists look for news tips in a specific topic for a news story when they don't have sources, they will, according to I1, ask their audience and community, "Do you know something about this topic?" or "Do you have a tip?" Every interviewee mentioned how compelling expertise and first-person experiences are: "*I always think that the most compelling comments are when people say, 'I worked in this industry for X amount of time,' or something that shows that they have specific experience or insight that might be more than other people ... in that topic*" (I1). As described by I9, personal experience can be searched for by using linguistic features, such as "I am" or "I have X" or "I work in something." When it comes to contributors with specific expertise, journalists can also "bookmark" information from these contributors based on their previous commenting histories: "*If the commenter has a good reputation for having comments... that would be really great [to find these people]*" (I1).

Community responses and trends. Many interviewees mentioned their desire to observe community responses and trends through quality conversations in UGC posts. I1 mentioned that a "*story with a lot of interesting comments about a topic that has already really engaged our readers ... and already getting a lot of quality conversation ... [is] a perfect identifier that this is going to be a good topic for us to continue the conversation and follow up.*" Though sometimes the volume of likes or replies could signal community interests and trends, interviewees argued that sometimes it is not enough to capture the quality of popular UGCs; instead, they look for quality conversations in response to specific UGCs. I7 indicated that the quality of replies has more power than the number of replies and thumbs-up: "*If somebody shares [writes] a story, it inspires other people to share [reply with] all kinds of other anecdotes, then that's how you know... you have hit on something people really want to read about, as opposed to people just being like, 'Yeah, I agree with your complaints.'*" This approach gauges how readers respond to previous news reporting and helps journalists better shape their future reporting. Long conversations mean "*something if your work is very focused on impact because I think the longer people keep talking the more they might be... motivated to act on something that affect[s] them*" (I4). Journalists do also monitor community responses via the number of people sharing articles. "*Facebook shares is our metrics of success because what we've generally found is that most reasonable debate on Facebook occurs on people's walls and Facebook pages because it's [about] our personal lives.*" (I3)

Questions. Questions raised by readers are an important mechanism for feedback as they help indicate whether a journalist might have presented something in a confusing way. As I7 reflected, "*I would say the first thing you want to do is just look for questions and see if there are any commonalities in what confused people or where people are all asking a follow-up question.*" Questions can be easily detected by looking for question marks and language like, "*I'm wondering about how this affects my neighborhood*" (I4). Questions raised by readers can be a good indicator for whether journalists should write a follow-up story and could help journalists shape their stories in a specific direction. I6 mentioned that one of their projects actually "*feature[s] stories in response to user-generated questions from the public.*" I4 also mentioned that "*if somebody has a further question about the issue, then that person can shape how it might approach another story or another follow up. I think it's really valuable, not even just for sourcing.*"

Different viewpoints and opinions. Different viewpoints and opinions can also be interesting for journalists to digest as part of their process. I4 mentioned there is a need for information when the community is talking about a particular angle to a story that they have not yet considered, such as "*That doesn't sound quite right to me,*" or "*That's not what I experienced.*" Journalists are also actively trying to reach other populations that are typically not easily reached. I3 mentioned, "*We're always having to think about how we can reach other populations that aren't as engaged with our content, especially in more rural communities.*" I5 added that "*an interesting point of view on something that no one else considered*" can be a useful UGC. I7 said that these different viewpoints and opinions could be detected by extreme sentiments, which could be quantified by sentiment analysis tools. In both starting new stories and writing follow-up stories, journalists don't necessarily fully understand the topic they plan to report on, and they can use UGCs as a way to get more knowledge and gauge public reactions to lead their future reporting.

4 PROTOTYPE INTERFACE AND INTERACTION DESIGN

In this section, we describe the design of a UGC sourcing tool. We first develop several scenarios based on the approaches and journalistic information needs that emerged in our findings reported in the last section. We then introduce several design goals based on the crystallization of our interview results, needs in our scenarios, and what we learned from prior systems and evaluations in this area. We illustrate and evaluate algorithms to support each design goal. Finally, we introduce the interaction and visual design of the interactive prototype to support these design goals.

4.1 Scenario Development

Based on the findings from our initial interview study (described in section 3), we identified several needs for journalistic sourcing using UGCs. Information needs include ways to find personal experiences and expertise, community responses and trends, questions, and different viewpoints and opinions. We also identified two overarching approaches to sourcing (i.e., *deep reporting* vs. *wide reporting*). These needs and approaches begin to articulate a design space for journalistic sourcing from UGCs. To better understand this design space and surface additional design requirements, we wrote three detailed scenarios that explored different combinations of needs and approaches emerging from our interviews [8]. The three detailed scenarios include: (1) finding personal stories and expertise for *wide reporting*, (2) finding community responses and trends, questions, and different/dissenting viewpoints for *wide reporting*, and (3) finding particular people of interest or with specific expertise for *deep reporting*.

This scenario-based design process helped us to better understand the context in which our prototype system might help journalists more effectively use UGCs in their sourcing practices. Ideally, a tool might be able to fully meet all journalistic information needs while also providing the ability to inform both focused and wide reporting approaches. However, here we decided to scope our effort towards supporting finding personal stories and viewpoints for wide reporting (scenarios #1 and #2) rather than finding particular people for deep reporting (scenario #3). This allows users to focus on the UGC content rather than the contributors of that content, and to evaluate the tool outside of an extended deployment that might be needed to understand the deep investigative process. We elaborate the final version of the scenarios in the context of a specific *New York Times* (NYT) article that shared U.S. immigrants stories of how they were told to "Go back to where you came from" [2]. These scenarios are visible in the prototype UI and serve to guide the evaluation procedure by setting a context for the participants in the study (see section 5).

4.1.1 *Scenario for Evaluation.*

The final scenarios focusing on the wide-reporting approach follow and were adapted for use in the evaluation:

After President Trump attacked four Democratic congresswomen of color in his tweets "*Why don't they go back and help fix the totally broken and crime infested places from which they came*", you, as a community editor, decided to write two community response articles by asking readers if they had been told to "*go back*". A huge number of responses flooded in and overwhelmed you, making it difficult to find the most representative and interesting responses. Therefore, you decided to try this system to help you organize all the responses and finish the two articles on time. The first article you are trying to write is an "investigation" article, in which you will focus on the specific angle of "*go back*" stories [Note: this information need is derived from Section 3.2.2], by looking at the community's discussions of this angle. For this article, you want to find as many interesting stories as possible with this angle from different people. The second article, on the other hand, is to present as many viewpoints as possible from the responses in the data, along with several representative responses for each viewpoint [Note: this information need is derived from Section 3.2.2 and Section 3.2.2]. You want to produce an article incorporating a range of experiences, summarizing as many of the perspectives from the community's discussions as possible.

4.2 Design Goals

Based on the insights of our previous study and the crystallization of user requirements and needs in our scenarios here we detail three design goals for our tool. We iteratively designed and developed our prototype to support these design goals (see Figure 1):

- **DG1. Rankings and Representativeness:** As discussed in Table 2, representativeness is essential to wide reporting, and we aim to support representativeness via rankings. Users should be able to see UGC posts ranked according to how representative they are of the posts in the dataset. This design goal is meant to help users better understand the gist of the community discussions, responses, or posts (See sections 3.2.1 and 3.2.2). We use a ranking UI since rankings are able to communicate a degree of relevance without filtering away information that may be of interest [16].
- **DG2. Interactive Re-ranking:** Users should be able to interact with the dataset by selecting and pivoting around posts of interest based on any particular sourcing needs they are pursuing. Once users indicate interest, such as by selecting posts in the dataset, or writing an example post to act as a query (similar to what journalists described in section

[2] https://www.nytimes.com/2019/07/19/reader-center/trump-go-back-stories.html

3.2.2), ranking of posts should be refreshed to make similar posts easier to find. This design goal is meant to support users in iteratively finding posts that reflect an interest in particular experiences, expertise, or questions (see sections 3.2.2 and 3.2.2).

- **DG3. Cluster and Diversity**: UGC posts should be clustered to support efficient browsing of different groupings of viewpoints to fulfill the needs mentioned in section 3.2.2. Users should also be able to select the data based on these groups and see summary labels to better support browsability. This design goal is meant to help users explore different angles from the dataset and support the wide-reporting scenario, which is our focus for this prototype (see section 3.2.1).

4.3 Algorithms to Support the Design Goals

Here we discuss how we designed the algorithms used to support each of the design goals in the prototype. UGC posts should be represented in a way that captures and makes comparable the semantic details of the posts. Averaging word embeddings is a simple yet powerful method to construct document embeddings and to capture meaning for every UGC post while reducing computational complexity [31, 44], and we decide to apply this approach to support our design goals.

4.3.1 DG1: Ranking and Representativeness. Users have the need to grasp all posts in a short period of time. To support this need, we must first define and then computationally operationalize *representativeness* so we can provide users with a ranking from most to least representative accordingly. In this work, we define representativeness simply as the most semantically "average" post from all posts, captured by the embedding space. To operationalize each post according to this definition, we use the distance between its own average word embedding and the centroid of the whole dataset's embedding vectors to find how average a post is. The more "average" a post is, the closer it should be to the centroid. Therefore, we define the default ranking as the distance between each post and the centroid, from closest to furthest (for a primer on these methods, see [63]).

4.3.2 DG2: Interactive Re-Ranking. To customize the ranking according to a user's inputs, such as a post provided as a query by example or a selection of posts identified as interesting, we define our re-ranking algorithm to sort the full list of posts and return a list of semantically similar posts compared to the input post, captured by the embedding space. We operationalize the re-ranking algorithm using the cosine similarity between the average embedding vector of users' inputs and each post's average word embedding vector. The higher the cosine similarity is, the more similar the post is compared to the users' inputs in terms of the semantics captured by the embedding space. Therefore, the re-ranking sorts all posts by the cosine similarities from the closest post to the furthest post based on the distance.

4.3.3 DG3. Cluster and Diversity. Users also need to understand the whole dataset from different perspectives and angles so they can understand the community's discussions more broadly. To support this goal, we implement a clustering function by using a K-means algorithm on the whole dataset's word embedding vector space and elbow methods to identify the optimal k from 2 to 10 automatically [38]. We label each cluster by using the top five words within every cluster (ignoring stop words and words that appear in more than 85% of posts or less than 2 posts) based on TF-IDF scores from each cluster so users could understand each cluster's key ideas. We also rank posts for each cluster from the most representative (i.e. closest to the centroid of the cluster) to the least representative when filtering the posts. However, once users provide inputs in the system, the system uses the re-ranking algorithm described in section 4.3.2.

4.4 Evaluation of UGC Representation in Word Embeddings

In order to support DG1, DG2, and DG3, UGC posts are represented using average word embedding methods. To select a performant method for our tool, here we report a comparison of different methods, including the average of pre-trained Word2Vec word embeddings (trained on Google news) [43], the average of pre-trained BERT embeddings (BERT BASE model, trained on Wikipedia and Book corpus dataset) [13], the average of Glove embeddings (trained on Wikipedia and Gigaword) [50], and Doc2Vec embeddings trained on a random sample of 100,000 NYT comments collected via their public API [3].

For the evaluation, we use a dataset from a callout survey used by the Washington Post in writing an article [4]. The dataset includes the full text of both published and unpublished responses but omits names, emails and other contact information (N = 472). The dataset was shared with the researchers for evaluation after the story was published. This dataset is particularly helpful because it encodes editorial evaluations of the posts: responses have already been clustered into semantic groupings (e.g., "*The Nationals are what made us Washingtonians*") and published if deemed sufficiently interesting by professional journalists. For each published response (N = 18), we compute a ranking based on the cosine similarity between that response as input and every other response in the dataset. In other words, we simulate the situation where each published response (which are known to be of editorial interest) is used to re-rank the rest of the dataset. Within each ranking, we assess precision based on whether responses come from the same semantic grouping (i.e. category) as defined by the journalists. We then calculate the mean average precision (MAP) score for each ranking to compare them.

We observe that BERT has the highest MAP score (MAP = 0.2284), followed by Doc2Vec (MAP = 0.2222) and Word2Vec (MAP = 0.2152). Glove has the lowest MAP score (MAP = 0.1488). Therefore, we decide to use BERT embeddings in our system. One thing to keep in mind is that there might be more newsworthy responses from the dataset that might not have been published as part of the article. These MAP results therefore reflect a lower bound on performance, but one which is nonetheless useful for comparing across embedding methods. We consider the implications of using these types of black-box models in the domain of journalism in our discussion.

[3] https://developer.nytimes.com/docs/community-api-product/1/overview
[4] https://www.washingtonpost.com/sports/2019/11/02/we-asked-nats-fans-how-it-feels-win-world-series-there-is-magic-world/

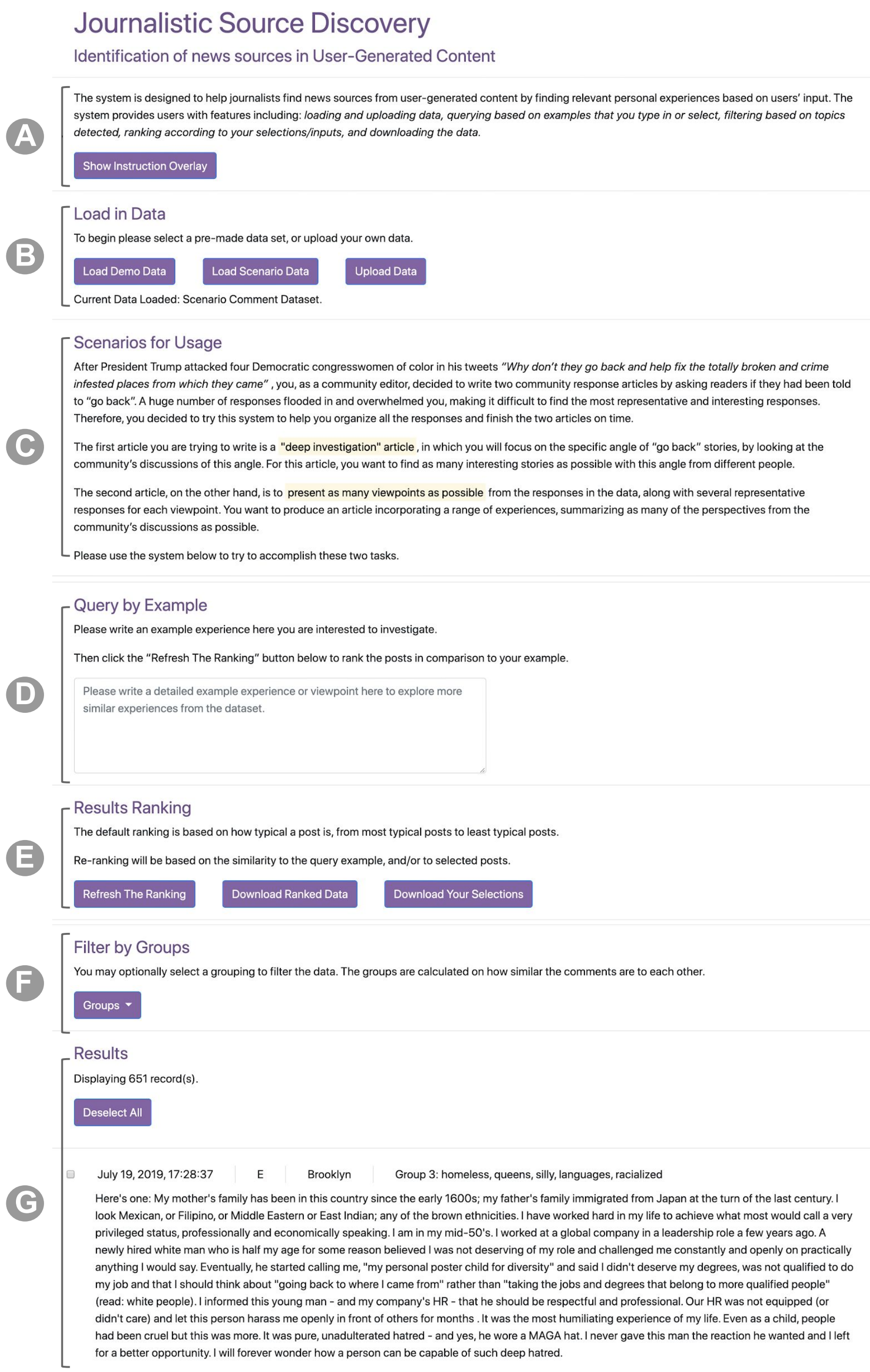

Figure 1: The prototype includes: (A) Overview and Instructions, (B) Load in Data: users can select demo/scenario data, (C) Scenarios for Usage, (D) Query by Example: users can provide an example experience or viewpoint to find more similar experiences or viewpoints by writing in the textbox, (E) Refresh Ranking: users can refresh the order of the dataset based on their selections and/or their example input, (F) Filter by Groups: users can choose to only browse certain groups in the dataset, and (G) Results: users can select comments by clicking the comments below and then refreshing the ranking based on similarities between selections and the rest of the dataset.

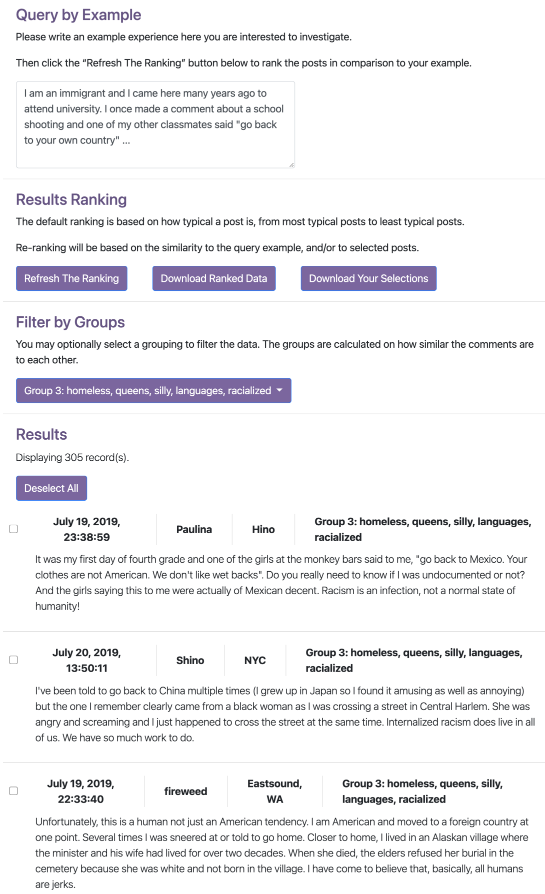

Figure 2: The ranking (top 3 posts) after providing example input in the Query by Example function and selecting Group 3 in Filter by Group function.

4.5 Interaction and Visual Design

The prototype interface (See Figure 1) has five key components: Load in Data (buttons), Filter by Groups (dropdown menu), Query by Example (textbox), Results Ranking (buttons), and Results (table). Users can choose different datasets to work with by clicking the demo or scenario dataset (more info described in 5.1), or upload their own dataset, following the same format as the demo and scenario datasets. The default ranking in the results table is ranked by our representativeness measurement (DG1) from most typical to least typical posts. Users can then filter the data to only view a certain group of posts by clicking that group in a dropdown menu. Each group in the dropdown menu is labeled by group number and the five top words by TF-IDF scores (DG3). The UI interface also shows the number of visible posts. Users can re-rank the dataset by writing an example experience/viewpoint in the Query by Example textbox and/or by selecting posts in the result table (DG2). For each post shown in the table, we show the published time, name of the contributor, location, and the group name and keywords as the metadata along with the post content itself, so users can get more context from the result table. Figure 2 shows a refreshed ranking (top 3 posts) with an example input as: "I am an immigrant and I came here many years ago to attend university. I once made a comment about a school shooting and one of my other classmates said 'go back to your own country'...", and with Group 3 selected. Users can keep refreshing the table by changing and/or adding selections and example queries in the textbox (DG2). Users are the one who make the final decisions about which posts are selected since the interface still show all posts unless they filter out data by groups. Once users are done with the selections and ranking in the table, they can click the "Download Ranked Data" button or "Download Your Selections" button to download their desired data and incorporate it into their broader workflows.

5 EXPLORATORY EVALUATION

In this section, we describe an exploratory evaluation of the prototype system with nine professional journalists to address RQ2. Participants were shown the system and then asked to work through a reporting scenario with the tool as they thought-aloud. The overarching goals of the study were to better understand how the tool and its design goals supported or otherwise impacted UGC sourcing practices, as well as to study the particular features of the system to evaluate their utility or shortcomings so as to inform future design. Our evaluation protocol was informed by the following criteria and underlying driving questions:

- **Sourcing Practices:** To what extent did users think the system supported their work practices in the scenario? How might it have better-supported journalists' practices in similar sourcing scenarios?
- **Relevance:** How relevant were the posts ranked by the system for users? Why? What other aspects of relevance might be helpful?
- **System Features:** Were there any issues users had in using the system that made it frustrating or difficult? Were there features that were particularly useful, or not useful?

Next we describe the datasets, procedure, and participants for the study, and then elaborate our findings based on a qualitative analysis.

5.1 Datasets

To support our prototype evaluation, we prepared one dataset for demo purposes and one dataset for participants to interact with in the context of specific reporting scenarios. Both datasets consist of comments from NYT news articles, collected using the NYT Community API [5]. For the demo dataset, we randomly select 17 comments from one NYT article, *Uber Hit With Cap as New York City Takes Lead in Crackdown* [6], to help demonstrate how to use the system to participants. The second dataset we prepared is used by participants in the main phase of the study. This dataset includes all available comments (N=631 at time of collection) for the same NYT article [7] that we used for scenario development (see section 4.1). The comments are organic community posts around "go back" stories, which serve as a meaningful corpus for participants to accomplish the tasks we ask them to pursue in the scenarios.

[5] https://developer.nytimes.com/docs/community-api-product/1/overview
[6] https://www.nytimes.com/2018/08/08/nyregion/uber-vote-city-council-cap.html
[7] https://www.nytimes.com/2019/07/19/reader-center/trump-go-back-stories.html

5.2 Study Participants

In order to evaluate the prototype, we recruited nine professional journalists who had experience working with UGCs to find news sources. Between May and August, 2020 potential participants were recruited to the study using purposive sampling, defined as the "deliberate seeking out of participants with particular characteristics" [45]. We recruited people by soliciting journalists on social media platforms (i.e. via Twitter and Google Forms), by emailing journalists who had participated in our interview study presented in Section 3, and by soliciting referrals from each participant. Three participants had taken part in our earlier interview study. Our participants' jobs range from community editor, engagement reporter, research analyst, journalist, and manager at newsrooms who routinely interact with and source different user-generated content (i.e. callout surveys, comments, social media posts) in their daily practices. Seven interviewees identified as female, and two as male. We include Table 3 for more information about each participant, including their role and organizational context.

5.3 Study Procedure

We first received participants' verbal consent at the beginning of each evaluation interview. Through online Zoom meetings and screen-sharing with participants, we first demonstrated how to use the system using the demo dataset to show users all the features in the system, including how to refresh the ranking based on their inputs in the textbox and/or selections in the table, how to filter the table based on different groups, what the default ranking means, etc. After the demonstration, we asked participants if they had any questions, and answered accordingly. We then asked the participants to read the scenarios we developed (visible in the UI, see Figure 1) and to then work on the tasks described in the scenarios so they could explore the prototype fully using the scenario dataset. The specific scenarios focused on the *wide reporting* approach identified in Section 3, asking participants to search for specific experiences related to "go back" stories and also explore and try to find different representative viewpoints. We observed how participants interacted with the prototype and asked them to say out loud what they were thinking as they interacted with the system [60]. After completing the tasks in the scenarios, we then engaged participants in a semi-structured interview, asking several questions related to the relevance of the ranking, the advantages and disadvantages of the system's features, and their overall assessment of the system. The sessions were recorded and transcribed for analysis using a thematic analysis approach [22, 23] that was quite similar to that described in section 3.1, incorporating constant comparison, analyst triangulation, and memoing while also conceptually reflecting the design features of the tool. Sessions lasted 44 to 67 minutes (Median = 55 minutes).

5.4 Findings

In this section we report our findings including high-level overall feedback, as well as more details on participants' practices with the tool, grouping and ranking features, and other suggestions made by participants to improve the tool.

5.4.1 Overall Feedback. Overall, participants were positive and encouraging about the system's features and potential uses, regardless of the size of their newsrooms.

Positive Feedback. P2 mentioned that the system overall was quite relevant and provided lots of new insights and angles that she could explore. P3 stated, "*I could totally see it being useful for the things I'm doing. I wish I could use it right now for various things I'm working on.*" P2 said that the tool could be a supplement to the other tools she has been using since she doesn't "*really see a lot of comment tools like this.*"

Negative Feedback. Participants also raised some critical feedback about the prototype design. An important piece of feedback from several participants was that it takes a while to learn how to use a system such as this including how functions work behind the user interface. But once they got the idea, it was easy to interact (P4). P1 stated that "*in order for this to be a tool used in the newsroom and for journalists, it would have to be very easy to use.*" More visual cues (i.e., color and bold labeling) could be helpful for journalists to better interact with systems like this (P8). It would also be worthwhile to include more instructions and explanations in the system and more training for users before their usage (P6). All participants asked some questions about the algorithms behind functions, e.g., how the system grouped the posts and ranked the posts, signaling a need for transparency on how algorithms in the tool work, an idea also supported by previous work on designing tools for journalists [16].

Another area of critical feedback relates to journalists' need to override or shape the algorithms. Participants were aware of biases underlying the algorithms and data supporting the tool. As a result, they wanted to override or shape the algorithms since they "*don't trust algorithms*" (P9). P6 added that machine learning can only straightforwardly evaluate the language but journalists have their own editorial standards that machine learning algorithms may not prioritize. P9 further suggested that designers ought to work with journalists to improve tool design since a lot of tools are not "*specifically built for journalism.*"

5.4.2 Overall Workflows. Based on the observation of their interaction with the system, some participants started by searching with their own queries directly, and some started by reading through the posts first. The differences in their workflows could be a result of their actual workflows in their sourcing practices, where some journalists focus more on finding exactly what they need from a dataset, other journalists are required to read them all in their actual daily routines, as P4, P5 and P6 stressed, or to double check if there are any mistakes from the system (P9). P4 specifically mentioned her urge to read them all, since she deals with callout data, and she doesn't want to miss anything from the audiences "*because people are taking the time to trust us with their stories,*" even though many of them might not be helpful. The urge to read them all, even with this tool, also signaled an awareness of potential biases brought by the tool.

If users were using the tool in their real sourcing practices, they might sometimes use the tool in a more exploratory way before developing search strategies:

Table 3: Participant Details

Participant	Role/Title	Type of Organization
P1	Community Editor	Large U.S. News Outlet 1
P2	Senior Investigative Researcher	Global Topical News Outlet
P3	Engagement Reporter	Large Non-Profit U.S. News Outlet
P4	Community Editor	Large U.S. News Outlet 2
P5	Engagement Reporter	Large Non-Profit U.S. News Outlet
P6	Journalist and Partner Manager	Large Non-Profit U.S. News Outlet
P7	Director of Growth	U.S. State Topical News
P8	Digital & Social Media Director	Large U.S. News Outlet 3
P9	Engagement Editor	U.S. Non-Profit Topical News Outlet

> If I was using this in my newsroom, my first step would be to examine the groups and to read some of the ranks, the top-ranked comments and then use that sorting because like I don't know what to search by, what to type out in that example experience. I feel like the writing and their experience might be more helpful if I'm like a little bit further along in a project where I read a lot of comments already and want to create like your own personalized search queries at the very beginning. (P4)

In general, users used the filter by group function to filter out things they were not interested in and used query by example and selections to find things they were interested in. When searching for a specific perspective (e.g., arguments from Trump supporters, experiences from women, etc.), all participants used the search by query function by writing an example viewpoint in the textbox to find more similar posts based on their inputs. When participants looked for interesting personal stories, some started with their own example inputs, while others just directly selected some posts in the dataset to search for similar relevant experiences. Once they found interesting posts, they used them to iteratively search for more relevant comments and/or expanded their previous example inputs in the textbox. They also downloaded their selections before they cleared the selections and started another round of searches to keep the selections locally on their computers. The tool also has the potential to explore different levels of groupings via downloading some groups as .csv files and then re-uploading them for further grouping analysis so that journalists can explore new groupings found by the grouping algorithm within a subset of their UGC collections (P9).

5.4.3 Groupings. Users used the grouping function to gauge the trends and different perspectives within the data, as we expected based on DG3. Users noted the number of posts under each group as a way to understand which group needed the most attention. Overall, users thought it was a "*really interesting component*" (P2) but needs some improvements to make it clear and to simplify it, especially for non-tech savvy users (P1).

Positive Feedback. Participants were able to figure out what content the groups reflect by reviewing top posts under each group. P3 stated it was helpful for journalists to quickly understand differences among groups through the group labels, even if they were not entirely accurate. Participants also used the grouping function to filter out posts they were not interested in to save time: "*It is really, really helpful, and cuts down a lot of time, especially in the first two groups where those are not even really that helpful.*" (P4). The function helps users cut down the volume they need to look at to go through all of the posts (P3 and P4).

Negative Feedback. Participants criticized that the keywords were not good labels for each group in our system. "*It's funny because like in this case, all the really interesting examples are grouped under a group three. And the keywords on it don't make any sense.*" (P3) P6 thought that "*it might not be worth displaying [keywords] because I find them confusing* " P1 mentioned, "*So I thought that the group's idea is definitely interesting. I'm just wondering if the way it's presented is clear enough.*"

Suggestions on Filtering and Grouping Design. Participants offered a good deal of feedback on the filter function. One key element is that they want to tag the groups manually or create their own groups (P3 and P7) to override the automatically generated groups. They also hope to choose multiple groups instead of only one group (P3) and to filter out certain groups (e.g., off-topic, overly political discussions, bad grammar and typos, short length) (P3, P4, P6 and P8). It would also be helpful for them to see an overview and breakdown of the relative size of the groups (P1, P7 and P8) so that it would be easier for them to locate where discussion is concentrated.

5.4.4 Ranking. Overall, participants agreed that "*it is truly interesting to be able to type like come up with a theory and have it pull up comments based on how close it can match*" (P1), and they were able to find "*some things that I didn't think that I would find [and] some new insights.*" (P2) Although some participants argued that the ranking itself might not save them time because they still need to read through them (P6), it is still a "*smart and sharp idea to find similarities or differences [in UGC posts].*" (P7)

Positive Feedback. The default ranking, sorted by representativeness from highest (i.e., the most semantically "average" post from all posts) to lowest, was found to be somewhat helpful by participants. P2 mentioned that the comment rankings in typical comment sections (i.e., ranked by published date) are not that helpful since the top-ranked comments are not great (newsworthy) for journalists, but she believed the ranking provided by our system could be beneficial to a lot of people who look for similar viewpoints and people to talk to.

For re-ranking based on users' example queries, P2 said, "*These comments came up that were in line with what I was trying to ask ... It really did switch when I did change my search terms from women to maybe GOP to actually naming like the Senate names; [the ranking] would shift based on my search queries.*" P5 thought the results from the example queries were "*super relevant*" and helped her "*start to notice themes as you are reading through, and then inform how you filter [and search].*" P5 further mentioned that the results from the refreshed ranking could inspire them to write a better example query.

Participants found the search by selection function to be very helpful and relevant, and "*actually more helpful than querying by example*" (P3). P3 elaborated this point, emphasizing that exploratory browsing might yield concrete examples that she could then use to re-rank and find other similar comments:

> I'm someone who would rather kind of read through the responses first to see what the different viewpoints are. And then...depending on the callout or depending on what I'm looking for, I would probably read through first and then do a lot of sorting by similarity to a particular comment, rather than doing a query by example.

Negative Feedback. P9 questioned whether the small data sample we provided and analyzed in the default ranking could afford representativeness due to the limited data. And some participants ignored the default ranking and started with their own queries and re-rankings (P2, P3, P8).

From our observations, for re-ranking based on users' example queries, participants tended to search by short queries, which didn't result in finding relevant comments (e.g., "*I support Trump*" in the textbox). And it took a while for them to come up with longer search queries to find more relevant comments (P4). Some participants (P1 and P3) mentioned that the ranking from their own queries was not that relevant. P1 mentioned that she couldn't find relevant posts by her own inputs. "*The results weren't quite as relevant as I had been hoping for. Maybe [if] I had more time and really played around with that, I would find something that was closer to what I was looking for.*"

The contrast in experiences among participants could be a result of their different example queries, and also the limitation of our scenario data, since the scenario data is from a left-leaning platform (i.e., *New York Times*), as mentioned by P2, and may therefore reflect the biases of the users on that platform (P9).

Many participants also mentioned that they used keyword searches a lot, which made it harder for them to understand the way this particular query function is implemented to find relevant comments based on the semantic relatedness of content rather than direct keyword matches. Also, as discussed by P4, it might not be that easy for participants to come up with a good example query without first reading some posts.

Suggestions on Ranking features. Participants mentioned several other dimensions of relevance that could be helpful when querying relevant content, including geographic, demographic, and political information (P1), sentiment (P1, P2, P3, P7 and P8), and recency (P6 and P8). Participants thought that it would be useful for them to have a keyword search in a UGC-sourcing system:

> I don't know if we switched everything over to a content relevant system how helpful keyword [search] would be, but I do think [keyword search] for some stories was good; let's say [you're] trying to figure out how many people died from COVID, seeing the word 'death' and being able to search by that in posts. I think it has ended up being kind of helpful for us. (P4)

5.4.5 Other Scenarios. Participants mentioned the system can be used for many different UGC formats, including news comments, form-based survey data (e.g., callout surveys), or different social media platform data such as comments from YouTube and Facebook. P4 specifically pointed out both short-term and long-term reporting can benefit from the system:

> If we were doing a breaking news kind of story and getting a lot of comments on something that we need to turn around very quickly, I could see that being great. I could see it being great if we were trying to put something together for social media based on some response that we had gotten. And then I can also see it being pretty helpful for like longer term investigations, too. If we... wanted to try to collect all responses from a questionnaire and group them by similarity, I can see it being really, really helpful for that.

Participants also mentioned that the system could help with their comment moderation (P1), finding misinformation and dissenting information (P2), writing readers' response articles (P3), brainstorming for new reporting ideas (P5), and shaping future investigation directions (P9).

5.4.6 Other design needs and suggestions. Apart from the feedback mentioned above, we also note several user needs and design suggestions that could be implemented in future systems:

- *Support for collaboration and integration:* As P3 discussed, in their real-life practices usually there are multiple people working on a big dataset. And she mentioned it would be useful for multiple users to "*be sorting the data together and tagging together.*" P5 discussed that it would be helpful for them to compare a couple different datasets since they deal with many callouts at the same time in their actual use-cases, and "*it is nice to be able to compare them across one another.*" P7 also talked about the need to connect with other platforms (e.g., google forms).
- *Trace conversations:* It would be helpful for users to be able to trace conversations across UGC posts (P8) so as to better maintain and see relationships between users' posts
- *Track progress:* P5 also mentioned that it would be helpful to have a tracker for users "*to see where I am in the responses*

so I know how many I have read through.", similar to what *GroundTruth*'s Grid offers [73].

6 DISCUSSION AND FUTURE WORK

Here we first consider the broader design learning and implications from our initial interview study and subsequent prototype evaluation, before turning towards a deeper look at how journalistic sensemaking practices related to UGC sourcing might be supported more fully. We then offer considerations related to the limitations of our work.

6.1 Understanding and Designing for Journalistic Sourcing from UGCs

In this work, we investigated journalistic sourcing practices, first in a qualitative interview study, and then by designing, building, and evaluating a prototype tool. Our initial qualitative interview study with journalists helps articulate a range of approaches and needs around journalistic sourcing from UGCs. In particular, our findings elaborate: (1) two UGC sourcing approaches which reflect different patterns of use and utility: *deep reporting* and *wide reporting*; and (2) a desire to tap into UGCs to identify personal experience and expertise, community responses and trends, questions, and different viewpoints and opinions. While commercial logics reflecting the need to drive traffic or engagement are present and subtly reflected in respondents' discussions of "time pressure" in leveraging UGC, the goals of the journalists we interviewed largely reflected deeper commitments to supporting or interacting with the community as a public service.

The evaluation of our prototype showed that our design was useful for journalists to find more similar UGC posts based on content relevance and explore the community's discussions from different angles. Participants were able to efficiently rank the UGC posts according to their selections and/or inputs to find more relevant content, and explore the dataset from their own perspective. The prototype appeared to encourage journalists to find their desired content more quickly, filter out unneeded content, and explore their data from unexpected angles. The evaluation confirmed that our design goals DG1, DG2 and DG3 were helpful in supporting journalists' actual sourcing practices. At the same time, feedback on our particular reification of DG2 and DG3 suggests there is further work to do in refining the expression of these goals in future systems. For instance, re-ranking based on content-based semantic similarity search (i.e. written input or selected examples) was found to be an interesting and useful model for interacting with the data; however, it needs additional research to reach its full potential (discussed further below). Still, participants expressed their actual need for a tool such as this in their real-life practices, offering inspiration for continued development of computational tools to support journalistic sourcing practices.

Our findings further reinforce the need for algorithmic transparency when designing for journalists, so that they can understand the algorithms shaping the curation of information and support their awareness of potential biases [14, 16]. A lack of transparency or trust could limit the uptake and reliance on such tools. Training and documentation is necessary to help journalists rely on the output of algorithms and use algorithm-based tools more efficiently. Building tools collaboratively with journalists, as P9 suggested, or participatory design more generally, may also be beneficial approaches so that journalists come to trust the results of such systems and see that their journalistic values are appropriately reflected [15]. Previous research has found that some news organizations have already worked with tech companies to ensure their goals are reflected in new technologies [2].

Our evaluation also showed some resistance from journalists to utilize algorithm-based sourcing tools in their sourcing practices since it is challenging to encode editorial interests using algorithms. Our prototype provides one simple way for incorporating editorial interests using a query by example method so that journalists could signal their expectations in the example queries. For other similar tools, we think it is important to provide flexibility and customization to journalists when designing tools for them since some participants mentioned their desire to override the algorithms, such as to create their own groups instead of using the automatically generated groupings, or to configure rankings based on customized weights as explored by *CommentIQ* [49]. Future work should consider how to mitigate journalists' concerns around algorithms, for instance by focusing on algorithmic transparency, while also incorporating flexibility and customization to meet specific editorial needs.

6.2 Supporting Journalistic Sensemaking Processes

Pirolli and Card [53] proposed an influential sensemaking model consisting of two key loops of activities: the *foraging loop* (i.e., seeking information, searching and filtering, and reading and extracting information) and the *sensemaking loop* (i.e., schematizing, building cases, hypothesis generation and evaluation, and presentation). We found that these loops of activity explain journalists' practices with our prototype fairly well, and journalists could enter into the loops at different starting points in terms of the prototype's functionalities. Some participants started with reading and filtering using the *groupings* function and *default ranking* function then *re-ranking based on users' example queries or selections* to build hypotheses around the UGC content. Other participants started with exploring their own hypotheses by *re-ranking based on users' example queries or selections* and then reading and filtering the UGC data using the *grouping* function. Our prototype demonstrated a capability to assist journalists in approaching the data from both starting points, by allowing journalists to enter the process from either *foraging* or *sensemaking* loops.

One possible reason behind different starting points could be related to the different types of UGCs with which participants deal in their daily practices. For instance, call-outs may already reflect scheme or subtle hypotheses in the way they are structured or questions are framed, with relatively smaller size and higher commitment from more targeted community members [47]. Journalists have a tendency to read them all, searching and filtering next, and then propose hypotheses at the end of data exploration. However, once journalists are faced with larger data sizes, less structured and less quality UGC content, such as social media posts, they tend to start with their own hypotheses rather than going back to information gathering in the *foraging loop*. Another possibility for different starting points may be driven by the context and goals of particular journalists. An investigative journalist going deep on a topic may have a pre-existing hypothesis to confirm (i.e deep reporting), whereas someone writing a round-up story of community reactions to an event needs to remain open to new information rather than constrained by a pre-existing hypothesis (i.e. wide reporting). We note that previous tools (e.g., *Vox Civitas* [18] and *GroundTruth* [73]) tend to primarily support one of the two loops in the sensemaking process. An interesting area for research would be to develop designs, perhaps similar to our prototype, that can be functional for multiple parts of the sensemaking model, so that journalists from different newsrooms with different practices could all benefit in their own way from the tool. In other words, because practices can vary greatly in terms of individual sensemaking needs, approaches, and goals, tools should be built with flexible entry points into the sensemaking loops.

In the next subsections, we discuss in more detail how several features of the prototype help with the foraging and sensemaking loops, as well as consider how future designs of journalistic tools could further support activities identified in the sensemaking model.

6.2.1 Foraging Loop and Grouping Function. The grouping function and default ranking in our prototype appeared to support journalists in their *foraging loop*, specifically in searching and filtering to extract and build schema. We believe similar grouping functions could further support the *foraging loop* by incorporating our participants' feedback, such as *keyword search* and *filters to exclude certain groups*. Our findings also suggest that there may be a need to more effectively support initial exploration. Some journalists have the need and urge to initially explore UGC posts before they make selections or define search parameters. This was despite our evaluation scenario offering one hypothesis that journalists could already investigate (i.e., "go back" stories). Sourcing tools might better incorporate ways for their users to explore a dataset by providing them different angles to look at in the dataset. This could be achieved through clustering methods similar to those used in our prototype, along with an iterative grouping function based on users' manual tagging. This could also involve developing additional newsworthiness detectors which help surface outliers, anomalies, or patterns that could serve as interesting entry points into the data [36].

6.2.2 Schematizing. The schematizing process in the sensemaking model reflects the exercise of organizing and representing evidence in some schematic way, which may for instance capture conceptual structure in the information space [53]. The grouping function in our prototype essentially operates as a schematizing component by providing journalists overviews of semantically related clusters in the UGC data. However, as participants mentioned, the labels of each grouping were not informative enough, and therefore the prototype failed to fully support the *schematizing* process. Future work should incorporate different labeling methods to help journalists better navigate and schematize the data. Research around cluster labeling promises several different approaches, such as summaries using cluster centroids [55] or cluster labeling enhancement using semantic resources such as Wikipedia and DBpedia [7, 29]. By utilizing different approaches, future research could explore what labeling techniques help journalists understand each cluster's meaning most effectively. One way to think of this is as a journalistically sensitive summarization function, which highlights the meaning of a cluster in a way that is most meaningful to journalists searching for angles based on groupings of posts. Some participants also suggested various methods to support schematizing, including providing an overview and breakdown of relative size of each group, along with visual presentations of sentiment analysis of each group. Integrating pre-developed scheme (e.g. pro vs. con, left vs. right, or others [48]) could further help initial exploration of the data. Research on how to support schematizing in journalistic tools is largely nascent [46] but we believe this is an essential and promising area for future development to support the connection between journalistic foraging and sensemaking loops.

6.2.3 Sensemaking Loop and Re-ranking Function. Our re-ranking by queries and selections function provides a way for journalists to build cases and create their own hypotheses in the *sensemaking loop*. Another possibility to help journalists search for support of hypotheses would be to bootstrap short queries (e.g., keywords as suggested by participants) into longer comments that serve as more robust inputs. For instance, longer example queries could be produced using synthetic text generation [79]. Journalists could first start with keywords to generate example query sentences, then revise the generated sentences provided by algorithms as the actual example query (i.e., hypothesis) to search relevant posts from the dataset. A different way to generate longer queries would be to utilize crowdsourcing to have people write hypothetical or fictional texts that journalists can then choose from. Journalists might then rank the real posts from their datasets using these fictional crowdsourced texts to define an average word embedding. Future work could explore how journalists interact with such synthetic or crowdsourced content, and whether such query paradigms would be helpful for journalists to understand UGCs from new perspectives.

6.3 Limitations

Though our prototype shows promising results in helping journalists in their UGC sourcing approaches, here we also note several limitations. The first limitation of the study is that we designed and evaluated our prototype around only two scenarios based on wide-reporting (i.e., focusing on the specific angle of "go back" stories and then presenting as many viewpoints as possible, described in section 4.1). This limits the potential usage of the tool for a broader usage, specifically for deep reporting. Future work should explore design opportunities around supporting a deep reporting approach.

Second, due to data availability (as well as broader privacy issues), we focus on finding UGCs to source from, rather than actually connecting and sourcing from UGCs. A deployment in collaboration with a professional news organization could open up new ways to integrate contact information such as email or phone numbers into the process. Moreover, connecting and monitoring social media platforms could also be interesting for researchers to explore, as it could unlock other social context for each post contributor.

The third limitation of the study is that we asked participants to evaluate the system in a one-hour session, which might constrain their ability to interact with the prototype in a more ecologically valid way. Participants' interactions with other tools (e.g. ScreenDoor and Hearken, which were mentioned) could also have a potential impact on their use of the prototype and where it fits within a broader workflow. This suggests a need for more detailed training and longer deployments with journalists. In the future, we hope to work with newsrooms over a longer period of time to understand users' thoughts and needs using their real datasets within actual workplace constraints [59].

Fourth, we think there are other technical approaches to implement content ranking and similarity, especially when journalists provide their own example input query. We were only able to evaluate and compare different embeddings using one call-out dataset provided by *The Washington Post*. Furthermore, there exist potential biases in pre-trained word embeddings (e.g., [3]). Future work should explore and evaluate different techniques to semantically rank data based on users' short queries, being sure to incorporate more professional input from journalists in order to ensure fidelity to editorial interests and algorithmic transparency to journalists. Such "editorially aware" design of ranking systems may also be a useful approach for ensuring that journalistic values are adequately reflected in computational tools [15, 32, 40].

Fifth, we chose not to focus on the differences across UGC platforms, but to focus on the commonalities of text-based audience materials from various platforms to design a tool to address broader cross-platform applicability. Future work might consider more specific tools that take advantage of platform-specific data types and availability.

7 CONCLUSION

In this work, we investigate journalists' sourcing needs and practices as they source UGC content, and design a prototype to support journalistic sourcing from UGCs. We articulate and contrast two UGC sourcing approaches: *deep reporting* and *wide reporting*, from our interview results with professional journalists. We furthermore report what journalists tend to look for in UGC posts, including personal experiences and expertise, community responses and trends, questions, and different viewpoints and opinions. We design a prototype for journalists to make sense of UGC posts and implement features for semantic ranking and grouping which support the tool's design goals. Our evaluation results indicate that professional journalists could benefit from the design for their UGC sourcing practices, while also offering useful avenues for future designs to explore. We discuss our evaluation results in the context of an information sensemaking model [53], which offers further opportunities for researchers to better design tools and algorithms for journalistic sourcing practices and beyond.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation Grant, Award IIS-1845460.

REFERENCES

[1] Tanja Aitamurto. 2015. Crowdsourcing as a Knowledge-Search Method in Digital Journalism. *Digital Journalism* 4, 2 (May 2015), 280–297.

[2] Charlie Beckett. 2019. New powers, new responsibilities: A global survey of journalism and artificial intelligence. https://blogs.lse.ac.uk/polis/2019/11/18/new-powers-new-responsibilities. Accessed: 2021-01-10.

[3] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc., Red Hook, NY, USA, 4349–4357. https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

[4] Shayne Bowman and Chris Willis. 2004. We media: How audiences are shaping the future of news and information. http://www.hypergene.net/wemedia/weblog.php. Accessed: 2021-01-10.

[5] Marcel Broersma and Todd Graham. 2012. Social media as beat: Tweets as a news source during the 2010 British and Dutch elections. *Journalism Practice* 6, 3 (2012), 403–419.

[6] Marcel Broersma and Todd Graham. 2013. Twitter as a news source: How Dutch and British newspapers used tweets in their news coverage, 2007–2011. *Journalism practice* 7, 4 (2013), 446–464.

[7] David Carmel, Haggai Roitman, and Naama Zwerdling. 2009. Enhancing cluster labeling using wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. Association for Computing Machinery, New York, NY, 139–146.

[8] John M Carroll. 2000. *Making use: scenario-based design of human-computer interactions*. MIT press, Cambridge, MA.

[9] Juliet Corbin and Anselm Strauss. 2007. *Basics of qualitative research: techniques and procedures for developing grounded theory* (3rd ed.). Sage Publications, Inc, Thousand Oaks, California.

[10] Stephanie Craft and Charles N Davis. 2016. *Principles of American journalism: An introduction*. Routledge, New York, NY.

[11] Dharma Dailey and Kate Starbird. 2014. Journalists as crowdsourcerers: Responding to crisis by reporting with a crowd. *Computer Supported Cooperative Work (CSCW)* 23, 4-6 (2014), 445–481.

[12] Leon Derczynski and Kalina Bontcheva. 2014. Spatio-temporal grounding of claims made on the web, in PHEME. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*. Association for Computational Linguistics, Reykjavik, Iceland, 65.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[14] Nicholas Diakopoulos. 2016. Accountability in Algorithmic Decision Making. *Commun. ACM* 59, 2 (Jan. 2016), 56 – 62. https://doi.org/10.1145/2844110

[15] Nicholas Diakopoulos. 2019. *Automating the news: How algorithms are rewriting the media*. Harvard University Press, Cambridge, MA.

[16] Nicholas Diakopoulos. 2020. Computational News Discovery: Towards Design Considerations for Editorial Orientation Algorithms in Journalism. *Digital Journalism* 8, 7 (2020), 1–23.

[17] Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. 2012. Finding and Assessing Social Media Information Sources in the Context of Journalism. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2451–2460. https://doi.org/10.1145/2207676.2208409

[18] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine. 2010. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *2010 IEEE Symposium on Visual Analytics Science and Technology*. IEEE Computer Society, Salt Lake City, UT, 115–122. https://doi.org/10.1109/VAST.2010.5652922

[19] Cristian Felix, Anshul Vikram Pandey, Enrico Bertini, Charles Ornstein, and Scott Klein. 2015. RevEx: Visual Investigative Journalism with A Million Healthcare Reviews. http://cj2015.brown.columbia.edu/papers/revex.pdf. Accessed: 2021-01-10.

[20] Herbert J Gans. 2004. *Deciding what's news: A study of CBS evening news, NBC nightly news, Newsweek, and Time*. Northwestern University Press, Evanston, IL.

[21] Andrew Thomas Garbett, Rob Comber, Paul Egglestone, Maxine Glancy, and Patrick Olivier. 2014. Finding "Real People": Trust and Diversity in the Interface between Professional and Citizen Journalists. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3015–3024. https://doi.org/10.1145/2556288.2557114

[22] Graham R Gibbs. 2018. *Analyzing qualitative data*. Vol. 6. Sage, Thousand Oaks, CA.

[23] Barney G Glaser and Anselm L Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge, New York, NY.

[24] Todd Graham and Scott Wright. 2015. A tale of two stories from "Below the Line" comment fields at the Guardian. *The International Journal of Press/Politics* 20, 3 (2015), 317–338.

[25] Juliette Harkin, Kevin Anderson, Libby Morgan, and Briar Smith. 2012. Deciphering user-generated content in transitional societies: A Syria coverage case study. https://repository.upenn.edu/cgcs_publications/20. Accessed: 2021-01-10.

[26] François Heinderyckx and Tim P Vos. 2016. Reformed gatekeeping. *CM: Communication and Media* 11, 36 (2016), 29–46.

[27] Alfred Hermida. 2013. # Journalism: Reconfiguring journalism research about Twitter, one tweet at a time. *Digital journalism* 1, 3 (2013), 295–313.

[28] Alfred Hermida, Seth C Lewis, and Rodrigo Zamith. 2014. Sourcing the Arab Spring: A Case Study of Andy Carvin's Sources on Twitter During the Tunisian and Egyptian Revolutions. *Journal of Computer-Mediated Communication* 19, 3 (March 2014), 479–499.

[29] Ioana Hulpus, Conor Hayes, Marcel Karnstedt, and Derek Greene. 2013. Unsupervised Graph-Based Topic Labelling Using Dbpedia. In *Proceedings of the sixth ACM international conference on Web search and data mining* (Rome, Italy) *(WSDM '13)*. Association for Computing Machinery, New York, NY, USA, 465–474. https://doi.org/10.1145/2433396.2433454

[30] Elihu Katz. 1957. The two-step flow of communication: An up-to-date report on an hypothesis. *Public opinion quarterly* 21, 1 (1957), 61–78.

[31] Tom Kenter and Maarten de Rijke. 2015. Short Text Similarity with Word Embeddings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (Melbourne, Australia) *(CIKM '15)*. Association for Computing Machinery, New York, NY, USA, 1411–1420. https://doi.org/10.1145/2806416.2806475

[32] Tomoko Komatsu, Marisela Gutierrez Lopez, Stephann Makri, Colin Porlezza, Glenda Cooper, Andrew MacFarlane, and Sondess Missaoui. 2020. AI Should Embody Our Values: Investigating Journalistic Values to Inform AI Technology Design. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society* (Tallinn, Estonia) *(NordiCHI '20)*. Association for Computing Machinery, New York, NY, USA, Article 11, 13 pages. https://doi.org/10.1145/3419249.3420105

[33] Sophie Lecheler and Sanne Kruikemeier. 2016. Re-evaluating journalistic routines in a digital age: A review of research on the use of online sources. *New media & society* 18, 1 (2016), 156–171.

[34] X. Liu, A. Nourbakhsh, Q. Li, S. Shah, R. Martin, and J. Duprey. 2017. Reuters tracer: Toward automated news production using large scale social media data. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, Boston, MA, 1483–1493. https://doi.org/10.1109/BigData.2017.8258082

[35] Josephine Lukito, Chris Wells, Yini Zhang, Larisa Doroshenko, Sang Jung Kim, Min-Hsin Su, and D Freelon. 2018. The Twitter exploit: How Russian propaganda infiltrated US news. https://mcrc.journalism.wisc.edu/files/2018/05/TwitterExploit.pdf. Accessed: 2021-01-10.

[36] Måns Magnusson, Jens Finnäs, and Leonard Wallentin. 2016. Finding the news lead in the data haystack: Automated local data journalism using crime data. https://journalism.stanford.edu/cj2016/files/Finding%20the%20news%20lead%20in%20the%20data%20haystack.pdf. Accessed: 2021-01-10.

[37] Neil Maiden, Konstantinos Zachos, Amanda Brown, George Brock, Lars Nyre, Aleksander Nygård Tonheim, Dimitris Apsotolou, and Jeremy Evans. 2018. Making the News: Digital Creativity Support for Journalists. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3174049

[38] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press, New York, NY.

[39] Brian Massey. 1998. Civic Journalism and Nonelite Sourcing: Making Routine Newswork of Community Connectedness. *Journalism and Mass Communication Quarterly* 75 (01 1998), 394–407. https://doi.org/10.1177/107769909807500213

[40] Melinda McClure Haughey, Meena Devii Muralikumar, Cameron A. Wood, and Kate Starbird. 2020. On the Misinformation Beat: Understanding the Work of Investigative Journalists Reporting on Problematic Information Online. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 133 (Oct. 2020), 22 pages. https://doi.org/10.1145/3415204

[41] Marcus Messner and Marcia Watson Distaso. 2008. The source cycle: How traditional media and weblogs use each other as sources. *Journalism studies* 9, 3 (2008), 447–463.

[42] Panagiotis Takas Metaxas, Samantha Finn, and Eni Mustafaraj. 2015. Using TwitterTrails.Com to Investigate Rumor Propagation. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work and Social Computing* (Vancouver, BC, Canada). Association for Computing Machinery, New York, NY, USA, 69–72. https://doi.org/10.1145/2685553.2702691

[43] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs.CL]

[44] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (Lake Tahoe, Nevada) *(NIPS'13)*. Curran Associates Inc., Red Hook, NY, USA, 3111–3119.

[45] Janice Morse. 2004. Purposive Sampling. In *The SAGE Encyclopedia of Social Science Research Methods*, Michael Lewis-Beck, Alan Bryman, and Tim Futing Liao (Eds.). Sage Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States of America .

[46] E. Motta, E. Daga, A. L. Opdahl, and B. Tessem. 2020. Analysis and Design of Computational News Angles. *IEEE Access* 8 (2020), 120613–120626. https://doi.org/10.1109/ACCESS.2020.3005513

[47] Mimi Onuoha, Jeanne Pinder, and Jan Schaffer. 2015. Guide to crowdsourcing. https://www.cjr.org/tow_center_reports/guide_to_crowdsourcing.php. Accessed: 2021-01-10.

[48] Deokgun Park, Seungyeon Kim, Jurim Lee, Jaegul Choo, Nicholas Diakopoulos, and Niklas Elmqvist. 2018. ConceptVector: Text Visual Analytics via Interactive Lexicon Building Using Word Embedding. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 361–370.

[49] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting Comment Moderators in Identifying High Quality Online News Comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1114–1125. https://doi.org/10.1145/2858036.2858389

[50] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[51] Chelsea Peterson-Salahuddin and Nicholas Diakopoulos. 2020. Negotiated autonomy: The role of social media algorithms in editorial decision making. *Media and Communication* 8, 3 (2020), 27–38.

[52] Swit Phuvipadawat and Tsuyoshi Murata. 2010. Breaking News Detection and Tracking in Twitter. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 03 (WI-IAT '10)*. IEEE Computer Society, USA, 120–123. https://doi.org/10.1109/WI-IAT.2010.205

[53] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. *Proceedings of international conference on intelligence analysis* 5 (2005), 2–4. https://analysis.mitre.org/proceedings/Final_Papers_Files/206_Camera_Ready_Paper.pdf

[54] Thomas Poell and Erik Borra. 2012. Twitter, YouTube, and Flickr as platforms of alternative journalism: The social media account of the 2010 Toronto G20 protests. *Journalism* 13, 6 (2012), 695–713.

[55] Dragomir R Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management* 40, 6 (2004), 919–938.

[56] Paul Resnick, Samuel Carton, Souneil Park, Yuncheng Shen, and Nicole Zeffer. 2014. Rumorlens: A system for analyzing the impact of rumors and corrections in social media. http://scarton.people.si.umich.edu/files/papers/Resnick%20et%20al.%20-%202014%20-%20RumorLens%20A%20System%20for%20Analyzing%20the%20Impact%20of%20Ru.pdf. Accessed: 2021-01-10.

[57] Md Main Uddin Rony, Mohammad Yousuf, and Naeemul Hassan. 2018. A Large-scale Study of Social Media Sources in News Articles. arXiv:1810.13078 [cs.SI]

[58] Raz Schwartz, Mor Naaman, and Rannie Teodoro. 2015. Editorial Algorithms: Using Social Media to Discover and Report Local News. *Proceedings of the International AAAI Conference on Web and Social Media* 9, 1 (Apr. 2015). https://ojs.aaai.org/index.php/ICWSM/article/view/14633

[59] Ben Shneiderman and Catherine Plaisant. 2006. Strategies for Evaluating Information Visualization Tools: Multi-Dimensional in-Depth Long-Term Case Studies. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization* (Venice, Italy) *(BELIV '06)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/1168149.1168158

[60] Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. 2016. *Designing the user interface: strategies for effective human-computer interaction*. Pearson, New York, NY.

[61] Pamela J Shoemaker and Stephen D Reese. 1996. *Mediating the message*. White Plains, NY: Longman, NY.

[62] Pamela J Shoemaker and Timothy Vos. 2009. *Gatekeeping theory*. Routledge, New York, NY.

[63] Adrien Sieg. 2018. Text Similarities: Estimate the degree of similarity between two texts. https://medium.com/@adriensieg/text-similarities-da019229c894. Accessed: 2021-01-10.

[64] Jane B Singer. 2010. Quality control: Perceived effects of user-generated content on newsroom norms, values and routines. *Journalism practice* 4, 2 (2010), 127–142.

[65] C Estelle Smith, Xinyi Wang, Raghav Pavan Karumur, and Haiyi Zhu. 2018. [Un] breaking News: Design Opportunities for Enhancing Collaboration in Scientific Media Production. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Montreal, Canada, 1–13.

[66] Einar Thorsen and Daniel Jackson. 2018. Seven characteristics defining online news formats: Towards a typology of online news and live blogs. *Digital Journalism* 6, 7 (2018), 847–868.

[67] Neil Thurman. 2018. Social media, surveillance, and news work: On the apps promising journalists a "crystal ball". *Digital Journalism* 6, 1 (2018), 76–97.

[68] Neil Thurman, Alessio Cornia, and Jessica Kunert. 2016. *Journalists in the UK*. Reuters Institute for the Study of Journalism, London.

[69] Neil Thurman and Aljosha Karim Schapals. 2016. *The contradictions of real-time online reporting*. Routledge, London, 283–292.

[70] Neil Thurman and Anna Walters. 2013. Live blogging–digital journalism's pivotal platform? A case study of the production, consumption, and form of live blogs at Guardian. co. uk. *Digital Journalism* 1, 1 (2013), 82–101.

[71] Peter Tolmie, Rob Procter, David William Randall, Mark Rouncefield, Christian Burger, Geraldine Wong Sak Hoi, Arkaitz Zubiaga, and Maria Liakata. 2017. Supporting the Use of User Generated Content in Journalistic Practice. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3632–3644. https://doi.org/10.1145/3025453.3025892

[72] Heli Väätäjä, Teija Vainio, Esa Sirkkunen, and Kari Salo. 2011. Crowdsourced News Reporting: Supporting News Content Creation with Mobile Phones. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services* (Stockholm, Sweden) *(MobileHCI '11)*. Association for Computing Machinery, New York, NY, USA, 435–444. https://doi.org/10.1145/2037373.2037438

[73] Sukrit Venkatagiri, Jacob Thebault-Spieker, Rachel Kohler, John Purviance, Rifat Sabbir Mansur, and Kurt Luther. 2019. GroundTruth: Augmenting Expert Image Geolocation with Crowdsourcing and Shared Representations. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 107 (Nov. 2019), 30 pages. https://doi.org/10.1145/3359209

[74] Gerret Von Nordheim, Karin Boczek, and Lars Koppers. 2018. Sourcing the Sources: An analysis of the use of Twitter and Facebook as a journalistic source over 10 years in The New York Times, The Guardian, and Süddeutsche Zeitung. *Digital Journalism* 6, 7 (2018), 807–828.

[75] Timothy Vos and François Heinderyckx. 2015. How gatekeeping still matters: Understanding media effects in an era of curated flows. In *Gatekeeping in transition*. Routledge, New York, NY, 39–58.

[76] Julian Wallace. 2018. Modelling contemporary gatekeeping: The rise of individuals, algorithms and platforms in digital news dissemination. *Digital Journalism* 6, 3 (2018), 274–293.

[77] Claire Wardle and Andrew Williams. 2010. Beyond user-generated content: a production study examining the ways in which UGC is used at the BBC. *Media, culture & society* 32, 5 (2010), 781–799.

[78] Andy Williams, Clare Wardle, and Karin Wahl-Jorgensen. 2011. *The Limits of audience participation*. Routledge, New York, 152–166.

[79] Ze Yang, Can Xu, Wei Wu, and Zhoujun Li. 2019. Read, Attend and Comment: A Deep Architecture for Automatic News Comment Generation. arXiv:1909.11974 [cs.CL]

[80] Arkaitz Zubiaga. 2019. Mining social media for newsgathering: A review. *Online Social Networks and Media* 13 (2019), 100049.