

BSR08_transcript_deidentified

SUMMARY KEYWORDS

research, data, people, twitter, researchers, tweets, context, social media, social media platform, analyzing, weibo, public health, individuals, publish, studies, users, questions, specific, instagram

SPEAKERS

BSR08, Sara Mannheimer

Sara Mannheimer 00:03

And so the summary of the research I sent this to you but is that so big social data, which I am calling it here, like social media, blogs, other online sources that people post, and archived qualitative data like interview transcripts, notebooks, diaries are similar. They're both qualitative data that people post online for one reason. And then they're used for different reasons for research or a new research. But I found that their respective communities of practice are under connected so big social researchers tend not to talk as much with qualitative researchers or people who do qualitative research. And so this research is trying to investigate how librarians can use data curation practices from each of these communities to bring them closer together for mutual benefit. To help qualitative researchers scale up their research by reusing data, and to help social media researchers be more responsible. And or just do more with social media data. Yeah. So any questions before we begin?

BSR08 01:49

Yeah, I've read your document. And I don't have specific questions. I may have other questions as we previously communicated after you interviewed me about more about your, for me to learn about your research and others, but I think I'll let you finish interviewing me first. And I hope I'm able to answer you questions from a specific perspective, which is [some details redacted], epidemiology, and digital health research. [Additional details redacted]. Digital health is actually a very, very broad domain. And it is a relatively new discipline, and when I say new—probably in the past 10 to 15 years. But if you consider public health as a discipline of over 200 years, this is really new. But digital health in itself contains a variety of components. And it ranges from telemedicine, to mobile health, to health communication analysis, using social media data, to using social media and other digital device data to predict epidemics, to using digital devices to help patients monitor their health and so on. So, the type of research that I have previously done primarily is analysis of social media data, initially using that as some of the variables that could help forecasting in epidemics. But it turns out, either perhaps because I'm not so good at forecasting or perhaps these, these data may not be a very good information input for for epidemic forecasting, and and my research took a turn, into the domain of digital health communication research. I also find that the majority of health communication researchers are not where they are, they're not comfortable with or they do not appreciate the type of more big data type of research in the communication realm. While there is a tiny fraction of communication researchers who do this type of big data analysis, they tend to form their own small community. It turns out that as I realized, because my training is not really in communication, but in epidemiology, so my concern is not really answering questions that communication scholars ask in terms of the types of theories and so on. And so that is why the type of research really doesn't really fit into the communication studies or journalism studies. So the reason why I spent five minutes talking about this is to... now you know the... it sets the background, sets the scene against which my answers to your following questions should be understood with that background.

Sara Mannheimer 05:12

Great. That was actually my first question was to tell me about the type of research you do and the type of data that you produce or use. So I feel like that really covers that. So is most of the data that you collect now from social media? Or do you do surveys and other types of data collection as well, for your epidemiology work?

BSR08 05:33

I may do surveys but I don't do much. This, I mean, even surveys, not really epidemiology when we talk of survey, usually, when you're talking about behavioral health science, or community health. Different names for the same thing like basically the psychological, sociological or economical variables that are predictors of certain types of behavior, for example, why somebody chooses to take the vaccine or not take the vaccine against COVID-19. But the type of... back to what I've said, the type of research that previous, previously that I've done, basically, prior to COVID-19. I will say it's not primarily quantitative type of research of social media, they are primarily Twitter. But I have done research in Weibo. I've also done a little bit of research in Instagram, and Pinterest. However, I, the type of research that I have done, I will say, is semi, semi quantitative, i.e., they're not very, they're not the type of thing that you hear from media where you have supercomputers doing deep learning and discover something that you really, we can't really consider but a computer algorithm generate that. I mean, I'm not a computer scientist on that. But at the same time, I'm not I was not doing the type of traditional qualitative research, where people are wanting focus groups, or this time one on one interviews, providing a lot of context to the specific tasks or specific documents or specific social media posts that they generate. Yeah, so meanwhile, right now, I'm more stepping back a little bit becoming sometimes why some, some, some form of an analytical support to communication scholars where they'd has to do more qualitative type of research. Yeah, so I don't know if I fully answered your question. But maybe, if I, you can let me elaborate, if you depending on what...

Sara Mannheimer 07:58

I think... I think that's good. I'm just looking for kind of a background as you said a foundation to help understand where the rest of the interview goes. So the next thing we want to do to establish foundation is to talk about a specific example, I'd asked you to identify a recent time when you collected big social data for your research, or use social data that was collected or shared by someone else. And this example, we don't have to stick to this example, alone the entire time, but it's helpful to have the specific time in mind as we walk through these questions. I'm going to be answering or asking questions about six key issues that I identified through a lit review, context, data quality and trustworthiness, data comparability, consent, privacy and intellectual property. And so we'll talk through any issues or challenges you encountered in those areas with this specific project. So do you have a project in mind that you want to use for today's interview?

BSR08 09:07

Well, any one of them. But I could choose. In fact, I want to, I want to talk about I want to use three different examples to illustrate. The one is a paper that I previously published in [an academic journal] about [infectious disease] misinformation. Okay, that was a comparison between Weibo data and Twitter data.

Sara Mannheimer 09:41

What's the first type of data We, Weibo?

BSR08 09:43

Weibo, Weibo. Yeah, Weibo, Weibo. Yes, that is the, that is China's equivalent to Twitter, because Twitter is banned in China as you know, and so W E I B O. Okay, so Weibo literally means microblog. Yeah, but but Weibo in itself is no longer is no longer something trendy in China. Most Chinese switch to WeChat. Anyway, so so it was in a historical context. But that is number one. And other examples. I think I have a study about using Instagram and Pinterest. That is more qualitative. I don't know if I have the link to that. Or you can find it on my website, I think.

Sara Mannheimer 10:42

Yeah.

BSR08 10:42

Yeah. So so that but I can briefly mentioned the challenge with regard to collecting information for Instagram, Pinterest, or any other thing other than Twitter. And then the third one is an ongoing study that my colleagues and I are doing using the Twitter data that Twitter released in 2019. Basically, they are tweets that were posted by accounts that were banned by Twitter. Okay, those accounts were considered by Twitter after the internal investigations as basically information operatives, by certain governments who are deliberately propagating this information or other whatever, whatever the names we call it, in order to achieve certain types of political goal, either internally or externally, i.e., involving international relations. So that a different context because it was questions about like, the context of privacy and that kind of thing. So I think this my using these three different examples can can give them more comprehensive answers to a question you asked.

Sara Mannheimer 12:06

Okay, nice. The sounds great. What data collection method do you use for each of these types of research?

BSR08 12:17

So for for Weibo, data, I think we were I've got a so so okay, so I collaborated with colleagues who had who previously used an API to call out Weibo data. They, my colleagues, my collaborators had, basically set up cohorts to follow certain publicly available Weibo accounts and to collate that data and then we were using that. So so it was an API. And for Twitter, I, if I remember correctly, it was also API, I think it was 1%. You could do either as a 1% API. So it was supposedly a random sample. But of the entire Twittersphere, even though we don't know exactly if it is really, truly random. Or you could use a certain keywords in order to search them. But there's also limits to how far you can get without paying. I have other studies that actually we have pay and buy a bigger share, i.e., for example, 10% stream or other for Twitter data. For the other studies that I mentioned, I think, for Instagram and Pinterest, I think Instagram, we were using API before they changed the user agreement. I think, after a certain—and I forgot at what time, Instagram changed the agreement, and severely limit the volumes of information that a researcher can download. So it basically make it extremely difficult to analyze. And I think for Pinterest, I think I had my students manually, manually screen for screen grab those information. And then the third study I mentioned for for the more recent one, the entire data set was released by Twitter as long as we provide a bonafide academic email address and our information then they will allow us to download the information for research.

Sara Mannheimer 14:41

Okay. Great. And do you have is this part of a grant funded project, for instance, that required like a data management plan? Do you have a specific plan for how you handle the data?

BSR08 14:54

For the studies that I have mentioned, they were not funded. Yeah, so so majority of my research are not grant funded.

Sara Mannheimer 15:04

And do, will you plan to release the data you collect it for any of these three projects? Did you publish it as supplementary material or published it anywhere?

BSR08 15:17

I did not. I do not. The major reason is that, especially, I mean, studying for the last example, I mean, the Twitter data is already released by Twitter for the public, for any researchers in the world, publicly available. For the other one, I think for, for Twitter data, whether we pay for it or we download it within the 1% stream that we're free to do. I don't think that we are allowed to upload those Tweets somewhere else, I think I think Twitter will require individuals researcher to directly get in touch with them and get permission to get hold of the Twitter data.

Sara Mannheimer 16:09

What I've seen people do is they are only allowed to publish the tweet ID. And then they have these tools that can happen is called Hydrate the tweet ID and bring in all the other metadata that was there. But then it means that if you hydrate the datasets, no tweets from people who have deleted their accounts or have deleted specific tweets, those tweets don't populate, you know. So that's what I've seen people do for Twitter. Yeah, it's not ideal.

BSR08 16:36

Yeah, sure. I mean, I did share that. I think one time, I mean, last year, somebody at another major university, I think someone, some students doing student projects with their professor, we request, to manually copy the data, of the Weibo and Twitter data for my [infectious disease] misinformation study. Because they want to, repeat the study for something similar, but want to compare with what I've done. And I sort of gave the deidentified a version of the manually culled tweets, and manually culled Weibo to the students in the other university.

Sara Mannheimer 17:20

Oh, that's so cool.

BSR08 17:21

So but that was between them and me through emails and and blah, blah, blah, I am not... So in essence, we are not releasing it into a public sphere, like, it is not in the certain database.

Sara Mannheimer 17:37

Right. Yeah. Did you see what they ended up doing with that data? Was it just a project? Or did they publish?

BSR08 17:45

Is a good idea? Maybe I should ask them. They did not contact me after that semester. So I don't.

Sara Mannheimer 17:52

That's so cool. Okay, let's move on to the first issue, which is context. So I have a little quote, to help understand where I'm coming from. And when I talk about context. So Halavais suggests that "when we collect data from social media platforms, just like when we collect data in traditional spaces, context matters. But the context of the social media posts could be absent or difficult to understand, since these

posts are, by nature, short pieces of text, or images or videos that are taken from a larger context of someone's personal or public life. And then the more data you have, the more the out of context effect can be compounded." So can you tell me about a time if there were any during the process of one of these or a few of these examples we have in mind when you considered the issue of maintaining or understanding the data's context?

BSR08 18:53

So to take the [infectious disease misinformation study] as an example, I think that there's... we are only looking at this very specific times where the [the outbreak was affecting both the U.S. and China] at the same time. So the context, are you referring to a specific outbreak at that time, is relatively clearly defined. And for the type of research that I've done previously, both the first example and the second example. Each of these are primarily are directly related to certain ongoing epidemics of emerging infectious diseases, and therefore, the contexts are relatively clearly defined. However, should digital health researchers choose to analyze data patterns to have conditions that are more prevalent... Take for example, I have another paper that was about [another infectious disease, centered around an event]. So my students looking at those tweets that were around the time of [the event]. You can say the context is clear in terms of we're looking at the conversations primarily promoted [around the time of the event]. But at the same time, the local context from which these individuals are advocating from, may not be clear to us as researchers, because the context of [the infectious disease] and the stigma associated with [the disease] in let's say, [example country] might be very different from from [a second example country]. And and given the fact that many individuals who use social media do not release or disclose their locations, we may not or we do not know where they are from, and therefore, the local context cannot be discerned.

Sara Mannheimer 21:06

So what can... does that concern you? Did you have like, did you talk about it in your paper? Or were there other things that you did to try to support more of this data collection, in context? Were you concerned that your results would that you wouldn't have the full picture? And if so, how did you talk about that?

BSR08 21:30

So I think context can be interpreted in multiple ways. Okay, so first of all, think about the geographical location. So I think it is, within computer science, a lot of people are talking about how to input missing data, or missing locations of Twitter users. And actually [I've seen entire research studies exploring new ways] of inputting the locations of Twitter users. So this is, could be very specific, very technical aspects of context. But then you can also have a different type of context, i.e., the context of the conversation. So somebody can be engaging in conversations about specific diseases, be they Ebola or Zika, or HIV or currently COVID-19. And when we talk about a context, you could talk about a local context, for example, I mean, the United States are more concerned about the response to COVID-19 by President Trump and President Biden, while someone else living in the United Kingdom will have a very different political environment, very different public policy, and therefore the entire conversations about COVID-19 on Twitter will be very different. But it could also be a different context, when we talk about thing about people's who debate about the safety and advocacy of COVID-19 vaccine on Twitter, as such debates can both be localized, but also international, because people are responding to the same vaccine, people are reacting to this the news of the same vaccine being introduced across the world. And so people across national boundaries across political jurisdiction, are engaging into a conversation about the same thing. And this in itself is a context for our conversation. So back to your question back to the examples that I have. I think that in the papers that I did my research, we do explain the challenges of locating these individuals, and we do not know where they are a lot of the

time, some of them we do have a geographical location, but many of them we do not. And therefore, it is very difficult to define whatever observations we have to a specific geopolitical jurisdiction, and therefore the health slash political slash economic context that these individuals may find themselves in.

Sara Mannheimer 23:58

Yeah. And so did you talk about this issue with anyone on your research team? Was it something that like you sort of as you were designing the research talked through and considered? Or are you thinking that? Yeah, tell me more.

BSR08 24:17

Well. Yes, we certainly do. And to use the third example that I mentioned: the Twitter data that have been released by Twitter, because it was really about different types of governments who are doing this information or slash whatever information operatives, what they were doing... the socio political economical context becomes very important. And so you know, either we are writing, certainly my communication scholars, colleagues, who are from different geographical political background, we certainly have to discuss the specific context in order to understand what the specific government alleged by Twitter who may or may not be doing information operatives? And if they are, if they were, what were their purpose what were their goals. And did they achieve anything? If so, what are they?

Sara Mannheimer 25:14

Yeah. Okay, great. Let's move on to data quality. Can you tell me a time during the process of one of the one or more of these research projects when you considered the issue of data quality? So for example, like missing data from the Twitter API, or but or bias among the Twitter users, or other issues of quality or trustworthiness of the data.

BSR08 25:43

Okay, so from a public health research point of view, especially from someone with an epidemiology background, a major concern that we have is whether or not the data are representative of the populations that we are interested in. And first of all, we know that people who use Twitter or any other social media, they are only a subset of the population, okay, they may not be representative of the general population. Now, for social media, such as YouTube, or Facebook, who may have more than half of Americans using them, they may be relatively representative in the context of the United States. But for social media like Twitter, which only a quarter of Americans are using it, it is already not representative among Americans. And when we're talking about geopolitical jurisdictions, in other parts of the world, the coverage, or the usage, in terms of the proportion of the population being active users of social media, they are not representative of the general population. So therefore, public health researchers usually have to indicate that whatever results they found can only be representative, among the... representative of the users who are social media active in their respective geopolitical jurisdiction, and not the general population. And that is why the it is extremely difficult to use social media data as an input for epidemic forecasting, because epidemics involve the entire population most of the time. While if you are talking about communication research, maybe the research question is different. You are interested in what people are talking about on social media such as Twitter, among those users who also are social media active. And you are not talking about the entire general population. And so that is why, in terms of epidemiology, I think the utility of social media data may not be as good as some advocates would like to admit. While for communication research, I think it is less a problem.

Sara Mannheimer 28:02

Yeah. So you basically only would ask certain research questions of Twitter... is what you're saying? You sort of use it for research questions that you think it will be more useful for?

BSR08 28:16

Yeah, and that applies to any social media.

Sara Mannheimer 28:19

Yeah. Again, thinking about, like how you work through, you know, like, you've come to all these conclusions about how to deal with missing data, or representative may miss of the data. How did you come to these conclusions, like did you consult resources or talk to your colleagues? Read other articles that people have been writing? What's been your process?

BSR08 28:48

Well certainly literature, literature review is certainly a big part of that. And I think we are very aware of the bias involved in social media data. I mean, it certainly became a major issue after some of my early papers. And I think for many individuals who more recently ventured into this field of trying to analyze the submitted data as if it represents somehow, an epidemic signal and something like that. It has been I think it is, it could have been an easy pitfall for many, until the time they realize that this is not really representative of population. So back to your question is, how would I be aware of this potential pitfall? I was a both literature reviews but also, reviewers' critique. I think I have submitted enough numbers of my manuscript and gotten rejected enough number of times before they get published. And so I think it has certainly gradually become a consensus among public health researchers, in fact, many public health researchers, especially people of the older generation, they are in general, very skeptical of this type of research.

Sara Mannheimer 30:16

Great. And so did you....Do you write a section in most of your papers that talks about this? Yeah.

BSR08 30:25

Limitations. Yes. Yeah. So certainly one, or more than one paragraphs of limitations and discussion, yes.

Sara Mannheimer 30:35

Great. All right. Let's move to data comparability. So during any of these examples, it seems like at least with that Weibo and Twitter one, did you combined multiple big social datasets to try to do your research?

BSR08 30:50

So for the two examples that I chose to talk about -- both the one with Weibo and Twitter and the other one with Instagram and Pinterest, certainly they are being presented within the same paper. But I'm not saying that we combined them, but we certainly compared them.

Sara Mannheimer 31:06

Okay.

BSR08 31:07

Yeah.

Sara Mannheimer 31:08

And what... Let's see, what strategies did you use to compare? Like, did you need to combine the data into one data set to do your analysis? Or did you do two separate analyses and then compare those two analyses?

BSR08 31:32

I think in... I think in both cases, we we, we would have certain types of framework, that that that generates the codebook that allows us to manually code the sampled data, so that we can... because in both cases, we are comparing content. And so we will need to have something to, something to compare with. So so basically, we need to standardize categories of contents to which we coded the social media posts. And then we can compare for example, whether or not this information about [the infectious disease] is higher in one data sets and the other on. There's no difference.

Sara Mannheimer 32:28

Okay, so the way that you... you used qualitative methods to analyze the data? Did you like, read through each of the tweets and coded them by hand?

BSR08 32:41

Yes, so for some of my studies, they were entirely qualitative manual coding. For some of these other studies we have used -- not the examples I chose to present here -- But for some other studies, we did use a Latent Dirichlet Allocation or some other type of topic models to allow basically the computers to identify potential latent, i.e., hidden topics. But then we still use manual coding to code the topics into humanly understandable categories of contents.

Sara Mannheimer 33:14

Got it. Okay, that's helpful. Great, did you encounter any other challenges? When you are trying to combine datasets in these studies?

BSR08 33:27

We'll take some examples, if you have... if you're doing comparisons involving data sets of different languages. So for example, I have another study that we basically compared the data between... on on the same topic, but it uses a different, different language. So we have different codebooks. Okay, so one of my most interesting studies, in fact, was comparing the same topic, but across different languages, tweets on Twitter. So even though it's the same platform, in fact, there are users of different communities who speak different languages. You would either need a coder who understands those languages, so you would need a team helping you. Or you would need to use technology, such as Google Translate API or Bing Translate or any other platform that may be free or they may charge charge money. But but either way, you will need help from human or technology to assist.

Sara Mannheimer 34:34

Yeah, great. Okay, let's move to informed consent. This is our fourth topic. So can you tell me about a time if any, during the process of these research projects when you considered informed consent?

BSR08 34:54

So technically, I don't have participants in the sense, we certainly have IRB approval, but they usually have certain types of exemptions. Because the type of data that we get are publicly available data. So somebody voluntarily consented to post those information online to let the world see it. And so we do not. So so. So IRB committed... did not consider that these studies require actually an informed consent, because technically, there are no participants.

Sara Mannheimer 35:33

When you're talking about sort of public health topics, has it become like, did you ever have any concerns about using the tweets? Or did it, was it pretty clear? Did it feel clear that if the IRB approved it and it's publicly available, these tweets are okay to use?

BSR08 35:55

I think so but of course, we I mean, we usually would not, in the paper publish, the Twitter handles, or the names of individuals, except for organizations. Let's say we can say the CDC or the the US government tweets about something. Or they are really prominent celebrities, which they can't I mean, claim, privacy in that context.

Sara Mannheimer 36:29

Yeah. And did you feel...? So you're kind of saying that you figured that individual people wouldn't necessarily expect their tweet to appear in a research paper, but bigger organizations and more public figures? You felt that those people would be okay with it? Is that right?

BSR08 36:55

It is just like, if you're reading a news platform by major media organizations, they would also, quote tweet, for example, and it could be unknown people, when most of the time they may say, oh, President, former President Biden's tweets say about something, or whether it is Bill Gates or the Gates Foundation's tweets about the importance of vaccines, for example, something like that. So I do not think that... I mean, so, so basically, for government or other major organizations, or celebrities, I mean, all politicians, I mean, they afford their publicly available social media posts, I do not think that privacy would be an issue. But but when we're talking about, there's some random people on the streets, and I mean, we may mention that content, but we usually will not write down we will not write their Twitter handles into the paper.

Sara Mannheimer 38:14

Yeah. Okay, this sounds good.

BSR08 38:19

The data we use are publicly available data. So anyone with internet access, technically, can join to it and a technically can hand weight any of these trades. So, so we are not analyzing any post that were made private by the users.

Sara Mannheimer 38:43

Okay. Nice. Okay, yeah. So the next topic is really closely related about privacy and confidentiality. Was there a time during the process when you considered issues of privacy? So you said you didn't use any tweets that weren't publicly available? What other considerations did you have regarding privacy of the users of Twitter or Instagram or Pinterest in some of this research that you did?

BSR08 39:17

I think I've already sort of explained that. I mean, yeah. So we, we, we analyze publicly available social media post, and so I don't see major uses of privacy. Yeah. And just like just like people, analyzing newspapers, I mean, we don't really have privacy.

Sara Mannheimer 39:40

Yeah. Yeah. So you're viewing the tweets as like, public websites that you're just scraping. Awesome. All right. Well, this is our last topic, intellectual property... Can you tell me about a time during your

research process when you considered intellectual property concerns. So could be like social media platform Terms of Service, or participant intellectual property, or your own intellectual property as a researcher?

BSR08 40:15

No, I didn't. I mean, technically as university researchers, whatever products I think we generate, I mean, the the... it goes to the universities. But I didn't I did not invent anything that can qualify as an IP, a patent. So yeah, no.

Sara Mannheimer 40:33

Yeah. Did you read the Terms of Service for like, it sounds like you did for Instagram and Pinterest and Twitter, and then take those into consideration as you were collecting the data? And did you try and meet the terms of service when you did?

BSR08 40:48

The paper that we published, the data were collected before Instagram changed their terms of service. So right now, I mean, after that, it really... after that, it is really impossible to do any meaningful research. And I do not see a way of getting around it. And so I did not do anything with Instagram after that. I did write a review paper with my students on... work about basically those image-based social media. But at the time, at the time, when I wrote that review paper, I think, if I remember correctly, in the review paper, actually saying that for the data that we as a team reviewed, that involve Instagram, they were done before Instagram changed their Terms of Service. And so those published, public health related social media research that involves Instagram, actually, they cannot be repeated in the future. Yeah, and I do not think... I do not find a solution to that. Unless, okay, individuals, say you've managed to recruit 500 participants, and then they consent, and then you we don't even download the Instagram photos. That's a different story.

Sara Mannheimer 42:20

Okay, so Instagram has said that there is no like, no more just like automated downloading, but you'd have to...

BSR08 42:27

Have they have a cap, they have a cap, which means that it is very difficult to do big data with and, and you want things to be representative. So that's the challenge. Right? So that is a major challenge in public health related social media research. Today, with the exceptions of those that that recruit participants in the first place and ask them for consent, for the researchers to read and download their social media posts... with the exception of those type of studies, most type of studies that the type of study that I have done before primary secondary data analysis, type of big data research, that do not involve directly recruiting participants for their concerns. Those type of research are heavily tilted towards Twitter research, because it is relatively easy. And Twitter's basically people consent to publish their, their own writing to the world, right? Because like things like Facebook, people use it to communicate with friends that they know in real life, while for Twitter, they are pretty much a platform to to say something to the entire world. The audiences are highly different. So because of that, so Twitter doesn't really have that much privacy issues as compared to getting data from Facebook or any other social media platform. And therefore you find 1000s of studies on Twitter. As Twitter data analysis, I'm talking about public health related one, I don't know anything not related to public health, but public health related social media research, primarily are Twitter data analyses. While we know that Twitter uses only a -- It's not very small, but it is not even half of social media users in the world. And therefore, it is not representative. Especially for younger people. They're using Instagram, using Snapchat, using

TikTok. While the majority of us who are professors are still analyzing Twitter, because we grew up with that. I mean, this is a social media platform of our generation, but they are not a social media platform of my students.

Sara Mannheimer 42:45

How do you like as a digital researcher, digital health researcher... What are your thoughts moving forward? Like, what do you think you'll do in the future? Do you think these Twitter studies are still going to be relevant? Or what are your plans for sort of addressing that, in your future research.

BSR08 45:27

I think that they, they are, they remain relevant, as long as a platform like Twitter or Facebook, has billions of users using them. We just have to be aware of its limitation. So so it may be a little bit cynical, or a little bit... it's half true and half a joke, that eventually, Facebook research and Twitter research is like doing research with older adults.

Sara Mannheimer 46:04

Right.

BSR08 46:05

Conducting health related research with older adults. Just like if we are doing research, about diabetes, hypertension, that sort of thing, because we are reaching that population group. But if we are thinking about doing, let's say, sexual health education, let's say substance abuse issues among young people, analyzing Facebook and Twitter would not help you with your research goal. And so I'm sure there are also researchers who are doing research with things like Instagram, Snapchat, TikTok, and so on, but I'm not so familiar with it. And I haven't ventured really into that substantially.

Sara Mannheimer 47:00

Does Twitter out of curiosity, does Twitter make available demographic information about its users? Or is that available anywhere?

BSR08 47:09

I don't think so. Most individuals do not... Most individuals who use Twitter, so, uh, the majority of them, do not disclose much of their own information. So technically, yes, they could. They could release their own information on their Twitter profiles, but the majority of them do not. And therefore, those who do is a tiny minority. And it is very likely they are not representative. They're not representative even of the Twitter users population, let alone the general population.

Sara Mannheimer 47:42

Right. Okay, cool. All right. So my last question is, are there any other issues or challenges that arose during these examples, these research projects that I haven't asked you about?

BSR08 47:56

I think that is a major issue here with regard to protecting privacy of users, which is very important as we are also social media users. And we, as users, we want privacy. I mean, if, if this issue has not been tackled, we may not choose to use social media platforms at all. But from the researcher point of view, this creates a hurdle that many researchers actually cannot overcome. And therefore, this is precisely the issue that makes public health related social media research heavily tilted towards Twitter. That means that while we as public health researchers understand the importance of the newer social media platforms as they are used by young people, we are not devoting as much time and resources into

analyzing them, and understanding the issues of the young people that deserve our attention. And I have not yet seen a major revolution in terms of how we as public health researchers are going to handle these issues. Yeah, because... because the way to go forward to overcome the issue of privacy actually would be to recruit a large enough sample of individuals, ask them for their consent, and then they have to surrender their social media data to the researchers.

Sara Mannheimer 49:39

Yeah. Okay, I was gonna ask if you have heard of people actually doing that. Yeah.

BSR08 49:48

Essentially, that could be... but you will need a big grant. And usually, it will be a certain subset of the population with whom you build a relationship. Let's say, let's say individuals who may get a grant from NIH to do research about delivering health information to a certain group of patients, maybe, let's say, HIV positive individuals or individuals with hypertension. And you're either using a social media platform as a health education platform... a channel that these people consent to participate, or you provide them with certain digital device that you promised them, this will help them manage their hypertension, and they sign their consent forms. They know what benefits they could get in return, and therefore they are willing to participate and provide you additional information on top of the demographic information that you collect -- maybe their social media information -- to assist with the research. But now we'll need NIH or other types of research grants that many [researchers] in tier two institutions or tier four institutions do not have. In fact the reality is, why so many people, including myself, are analyzing social media data in the first place, is because we do not have a big grant to recruit 1000 people.

Sara Mannheimer 51:12

Right.

BSR08 51:13

For research, right? That is why the convenience of getting secondary data is that yeah, it is, even if we pay for it is still not too expensive. We have expertise analyzing them. And they have okay, we can get papers published and satisfy our own curiosity. And yet meet the needs of the University and people move from professor to associate professor.

Sara Mannheimer 51:38

Yeah, that's such a good point. And I feel like that's part of the argument for sharing qualitative data, too, is like, well, it's, you know, it's time consuming. And I have, you know, I had to ask you for this huge favor to spend an hour of your time talking to me, you know, to gather qualitative data. And so if we can publish more than there will be just like with social media, it will be like there is this, you know, this true treasure trove of information that then researchers can use to ask different questions. So, yeah, that's such a good point.

BSR08 52:16

And also another question was, the pace of academic research is slow, relative to the development of social media platforms. So take the environment of social media in China as an example. [A few details redacted regarding BSR08's research experience]. Because my collaborator collected those Weibo data sets for other research topic. And I simply... we use them for health, looking at health keywords and health-related questions. By the time some of my early papers got published. I think there's a cohort of researchers of similar age group who picked up the ideas, and they followed my ideas and repeated that for many other health topics. In the following two, three years, I moved on to analyzing Twitter data. However, by the time many of these research got published, and get the attention of the

non Chinese speaking, scientific community, let's say, the western scientific community, the use of Weibo, among Chinese speaking individuals has significantly declined. Or they they remain users, but they become inactive. That was because of the Chinese government's policy of suppressing voices of dissent within the social media platform, and also with the emergence of WeChat, which is pretty much China's equivalent of WhatsApp, with additional functionalities. And the Chinese public basically switched over to WeChat. And likewise, by the time public health researchers realize the importance of WeChat and start to find ways to analyze WeChat -- which is actually very difficult -- the, the WeChat sphere, so to speak, in China, and the Chinese speaking individuals around the world has also evolved. Primarily because of the major suppression of voices of dissent on WeChat by the Chinese government in the recent two years. So the social media landscape changes a lot, very quickly, but the cycle of academic productivity has longer wavelengths so to speak.

Sara Mannheimer 55:23

Yeah. That's such a good point. Yeah, I mean, I'm finding that even in the three years that I've been working on my PhD, it's like each year that we you know, the Twitter Terms of Service change, or you know, Instagram's Terms of Service, or yeah, people are moving platforms. I've definitely experienced that, too.

BSR08 55:42

So back to your question, I actually want to say one more comment is that I see your I see your perspective, as, as a librarian and Data Manager, how you could facilitate sharing of data between researchers and prevent duplicating effort and so on. But by the time a data sets can be shared, it's old.

Sara Mannheimer 56:02

Yeah.

BSR08 56:02

So it's a historical data set. So so for the fast-changing field that I am in, let's say we are researchers of emerging infectious diseases. We are all analyzing COVID related data. Ebola of 2014 sounds like ages ago. Nobody's talking about Ebola right now. Nobody's talking about Zika right now. So yeah, yes. Even if I am available to or I am capable of sharing data, that was six years ago. Who is going to use it? I mean, maybe a medical historian, but not the active researchers who are chasing the next epidemic.

Sara Mannheimer 56:43

Mm hmm. That's a really good point. Yeah. I mean, I think, you know, you could use it for longitudinal studies, or comparative studies where you look at how, you know, Zika spread or, you know, in comparison to a new similar disease or something like that. But you're right. It's like, if... thinking about science as this constant pursuit of the cutting edge, that's problematic. Yeah. Cool. All right. Well, that's all I have for you.