BSR09_transcript_deidentified

SUMMARY KEYWORDS
tweets, data, research, twitter, collected, people, infrastructure, privacy, intellectual property, post, issues, users, context, thinking, paper, news articles, area, api, question, information

SPEAKERS
BSR09, Sara Mannheimer

Sara Mannheimer 00:03
The research basically is about connecting two communities of practice that I've found to be under-connected, which are people who do research with big data, big social media data, which I'm calling big social data in my project, and then people who reuse or share their qualitative data. So I'm sort of envisioning social media data as a type of qualitative data. It's text, images, videos that's shared online. And then when you use it for research, it's kind of like you're reusing it, you're putting it to a new type of use than what it was originally intended for. So my hope is that by talking to researchers who have done big, big data research, and by talking to qualitative researchers, and talking to data curators like myself, I'll be able to develop a set of strategies that data curators and librarians can use to support responsible research in both of these areas. So that's why I'm talking to you today. I think one of my other research participants recommended you, saying that you were working with this type of data and that you might be good to talk to so. So yeah, the interview is going to be structured around these six issues that I've identified, which are context, data quality, and trustworthiness, data comparability, consent, privacy and intellectual property. So I'll ask six questions. I've got a little intro and conclusion question. It should take about an hour, maybe less, sometimes 45 minutes. Yeah. So my first question is, tell me about what type of research do you do? What type of data do you generally produce or collect?

BSR09 02:07
Sure, yeah. I'm a postdoctoral researcher at [a university]. So I'm in [an engineering department]. So what we are what data we are working on is about like, the operation or the performance of infrastructure and environmental systems. So yeah, also, the scale of the data are at a city level. So we not only focus on the infrastructure, but also we look into the activities of the population because you know, the urban systems are always composed of humans and infrastructure. So the interactions between them are very important for us to understand the performance of the infrastructure or the the improvements that we needed to make to improve the infrastructure for future needs. Right. So yeah. Also, yeah, including social data, we, we usually collect data from the physical world, and also the virtual environment, like on a platform like Twitter or another social media platform. And for physical world data, we're usually looking to the human activities. So the like... damages in the road networks. So those are collected from like sensors or, like, human human powers are used to detect those kinds of damages on road.

Sara Mannheimer 03:41
Yeah. Great. Okay, so I had asked you, I think I sent you an email a couple of few. It was a few weeks ago that we first interacted, but I asked you to identify a particular example of a research project you've

done recently when you collected big social data, or reused social data that was collected by someone else. Do you have a project in mind?

BSR09 04:08
Yes, we have a recently we have a project related to using social media data on Twitter to detect infrastructure disruptions. So the case is [a natural disaster in a U.S. city]. So yeah, we collect like about how many social media... maybe 20 million or 10 to 20 million Twitter data. And yeah, based on these data, we apply some natural language processing techniques to identify the content which are always supposed... which are related to the specific infrastructure in [the U.S. city], or [in city water sources]. We use... and also some specific roads or highways, toll roads, etc. and hospitals, airports. Yeah, a lot of infrastructure can be... the performance of this infrastructure can be assessed or sensed using this social media data.

Sara Mannheimer 05:21
Okay. Um, tell me about your data collection method. Did you use an API?

BSR09 05:27
Yes. We used a power track API provided by Twitter.

Sara Mannheimer 05:35
Is that one that you purchased? That [your university] purchased?

BSR09 05:38
Yeah, we have to purchase.

Sara Mannheimer 05:39
Yeah. Okay.

BSR09 05:40
They have the free API. The free streaming API. But, you know, the data is not complete if we use the streaming API. So we have to use the power track API.

Sara Mannheimer 05:53
Yeah. Is the power track API complete? Or does it just provide more tweets? Are they talking...

BSR09 06:00
It's complete.

Sara Mannheimer 06:01
Okay, yeah. Nice. Great. And then was your project grant funded? Like? Did it require certain treatment of the data? Did you have a data management plan or any way that you are required to treat to handle the data?

BSR09 06:18
Um, I'm not sure the funding sources for this data. But yeah, we definitely have like a, have a plan, or we have this actually, we have the server and we have the data storage account [specifics of data storage services redacted]. We use that to manage the data or share the data with teams. And also we use that platform to do the data computation, computational experiments.

Sara Mannheimer 06:48
Okay. And then, last question about the data. Do you plan to share the data beyond your research team, the Twitter data?

BSR09 07:01
Oh, the data is not allowed to be shared. Yeah, we have to keep data confidential.

Sara Mannheimer 07:08
Okay. And is that an agreement that you made? I guess we'll talk about this more throughout the interview. But is that an agreement you made with Twitter?

BSR09 07:15
Yes.

Sara Mannheimer 07:16
Okay. Through... because of the API? Because I have heard that you're allowed to publish just the tweet IDs. So just like the ID number?

BSR09 07:26
Um, I think that since the data is... since we purchased this data, so we are not even allowed to share the ID.

Sara Mannheimer 07:35
Okay. And do you plan, how long do you plan to store and retain the data? Do you know?

BSR09 07:48
We can store the data for, for us, I think.

Sara Mannheimer 07:54
Yeah, yeah. And do you think you will store it like, will it continue to be useful to you and your team?

BSR09 08:00
I think we should probably hold on to this data for five years. And after five years, we may, because [after it has been] five years, I think no one would be interested, that we'll definitely switch our research directions or topics to other disasters or other contexts.

Sara Mannheimer 08:28
Yeah. Great. Okay, so I'll start with my first of the six issues context. So I have a little quote to help sort of communicate what I mean by context. It's been suggested that "when data is collected from social media platforms, just like when you collect data from people in traditional spaces, context is important. However, the context of a social media post could be absent or difficult to understand, because social media posts are by nature, short pieces of text, or images or videos that are taken from a larger context of a person's personal and public life. And then this out of context effect could be compounded when data are masked at this very large scale, millions of tweets." So can you tell me about a time during this research when you considered the issue of maintaining and understanding the context of the tweets that you collected? So like contextual information about the community where the data was collected or about the respondents or about the situation they were in, for instance.

BSR09 09:42

First, I think when we start to collect the data, we are usually using several filters like keywords that related to [the natural disaster] or the the geotags in the tweets, which can help us to locate the tweets posted by the people, in [this U.S. city]. So, this helped us a lot to understand the use of tweets posted or related to the context that we are trying to study. Also, we use natural language processing approaches to, to filter out some tweets that does not make any sense, where there are some tweets only, like a post, punctuations or emojis, or some tweets do not have any, like contextual information, or like other keywords that can make sense to us. So, we remove those tweets, so we only focus on tweets that can deliver effectively information to us related to the disaster.

Sara Mannheimer 10:53
Did you have any, like as a research team, as you were working through your research design, did you have any concerns about context about how these what the meaning of these tweets would be and how you sort of pulled them together to create a bigger picture of what people are thinking during [the disaster] that you can connect with your infrastructural data?

BSR09 11:25
I think in our research, there will always be concern about the context. Because you know, even some contexts that use... you may think it's related to the project or related to other problems that we are trying to solve. But maybe it is posted for other cases or other contexts or other events. So it is... it will always be a concern. But we use the filters to try to reduce the uncertainties from this kind of perspective.

Sara Mannheimer 12:08
Did you consult with anyone... like what resources did you consult, or people, to come up with this strategy that you ended up using of like filtering in order to support context?

BSR09 12:22
To support context, we have to look at the inverse effect the area of the events like [this natural disaster] or a more recent event, which is [another natural disaster], right. And also, from the results we will also compare this with our results generated from the Twitter data and also our model with information collected from news articles. Because usually news articles are confidential, or not confidential, are trustable. Right. So, you know, we use information from news articles and government reports to validate the information we gathered from Twitter.

Sara Mannheimer 13:06
Oh, interesting. Okay, we're going to talk more about comparing different data sets too. So actually, let's just talk about that now. So during this research, did you compare or combine multiple datasets? Like when you used the news articles was that a data set that that you then computationally compared with the tweets?

BSR09 13:29
Actually, we do not do the quantitative comparison. We just use news articles to validate the results we obtained from Twitter. For example, we use the Twitter data to map the timeline of events in a specific neighborhood, you see, [when the natural disaster is more or less impactful], when the power is out, and when the power is recovered. So we use the Twitter information to to to to map the timeline of the events in the neighborhood, then we correct the correct news article to find the EFCC information are consistent with the information posted in the news article.

Sara Mannheimer 14:13

Oh interesting. Did you do any other type of combining of data in this project?

BSR09 14:20
Um, I think we only use the Twitter data. So we do not do any combination.

Sara Mannheimer 14:27
Have you ever done that in any project?

BSR09 14:34
In the past project, if we want to combine data, we only combine data which are compensated to each other. For example, we have the Twitter data and also we have like a point of interest where the data is used, and then we can integrate them to look at where the tweets were posted around the which area and [what damages are occurring in that place]. Also, we may compare it with combined with social demographic data to see which area are populated, or are more concerned by the user on Twitter, but we did not turn by any other similar data sets together.

Sara Mannheimer 15:20
Yeah, okay. Did you.... Okay, I see. Sounds good. Okay, let's move on to data quality, and trustworthiness. So can you tell me about a time during this research when you consider the issue of data quality? For example, missing data, or bots, or bias?

BSR09 15:46
Yeah, we considered the bots on Twitter actually. So to say, there might be some bots, but we may not be able to identify them manually. So there is a tool available online, which is publicly free tool. And if we only want to test a few tweets, it should be free. And if we want to test millions of tweets, we have to pay for something. So us-, using this kind of tool, we can detect the accounts which might be the bots, and then we remove the tweets posted by these accounts.

Sara Mannheimer 16:32
Okay. What other did any other data quality issues arise during the research?

BSR09 16:44
Another issue might have been geotags on the tweets, you know, usually there's a, around like one or 2% of tweets are geotagged. So if we want to assess the area based on the... or assess the damages of the area based on geo tagged tweets, it will definitely have a lot of like quality issues, because usually people don't like to post the tweets with the geotags. So so to do to, to like overcome these issues, we developed a method to extract the locations from the content of the tweets, then we locate these tweets in the specific area, which is a mentioned as an address. So we can use this to locate the the tweets in the area, and they use that tweets to assess the damages in the area.

Sara Mannheimer 17:42
Were there enough tweets? I mean, I guess you did collect millions, but were there enough that you were able to extract a location from enough of them?

BSR09 17:51
I think, is still not sufficient. Yeah, but it's improved. A lot of which is only based on the geotag tweets.

Sara Mannheimer 18:03

Yeah. Um, and did you... thinking about data quality. Like, did you consult with anyone on this? Did your team talk together about these issues? And yeah, tell me more about what the process was?

BSR09 18:21
Sure. Yeah, we usually discuss with our own team members. And we also look at the literature, which has been published before. But actually, previous literature does not deal with these issues very well. So we have to come up with our own ideas, or approaches to, to somehow overcome this issue. Yeah.

Sara Mannheimer 18:49
Yeah. And then, how did you like communicate these strategies that you had used? Did you write about them in your paper or otherwise sort of document these strategies that you use to support data quality?

BSR09 19:07
Yeah, we always describe this approach clearly in our published papers.

Sara Mannheimer 19:14
Yeah. Okay. Let's move to informed consent. So can you tell me about a time during this research, if any that you thought about informed consent of the participants, Twitter users?

BSR09 19:38
Oh, could you repeat that, sorry.

Sara Mannheimer 19:40
Yeah. Was there a time during this research where you thought about the idea of these users agreeing to be part of the research? consenting to be part of your research?

BSR09 19:52
Because, you know, Twitter has, like a consent agreement with users. So as long as they sell the data to us, we can use that. So we didn't have any concerns about this.

Sara Mannheimer 20:07
Yeah, it's interesting that you... that Twitter sold the data to you. And so it was, you were actually sort of making an agreement with Twitter to, to download the data. And then did you talk about that at all with Twitter? Or did your research team? Or was it just we have permission to give these tweets to you, and therefore...?

BSR09 20:32
Twitter has agreement with the users. So they have this information on their website. And then when we purchase data, we are also assigned a sign a very strict agreement. They mentioned that this kind of agreement with our users and with us, so yeah, we talked about that.

Sara Mannheimer 20:55
Yeah. Did you do you as a researcher, like, have any questions about that? Or did it seem like that made sense? And you thought that the participants would be okay with the research that you're doing?

BSR09 21:13
Yeah, we actually, we thinking about that. And also, we submitted our IRB application to get approved. So yeah, we... any publications are based on the approval of IRB. So yeah, we definitely think, are thinking about these things, thinking about that, and we got approved. And yeah.

Sara Mannheimer 21:39
Okay, sounds good. How about privacy, and confidentiality. So this is the fifth out of six. Can you tell me about a time during this research, when you considered issues of privacy for you know, like protecting the data, or protecting the data during your research? or thinking about the people whose tweets you're pulling? Like, did you publish any tweets in the paper? any full text tweets or...? What issues arose as you thought about privacy of Twitter users?

BSR09 22:20
Okay, I think there are two issues we always keep in mind. The first issue is user name, or user ID information. Also, also, although we analyze the tweets using like a social network analysis, but we never show any IDs in our publications. I think, as you see, it's very important to protect, like the privacy of the users. And another issue is, I think it's the, like content in the, in the picture. Sometimes people may have a face in the picture. And we, we don't want to show these in that, in our paper. So we usually remove any of the results or any of the, like, [any data] that contained a photo of any people. So this is what we... what actions we take to to protect privacy.

Sara Mannheimer 23:23
Is that like a policy in your lab? Like, did you... Is that something that... just wondering, like how you came up with that? You know, thinking about how Twitter sold you the data, and said all these people had consented for their data to be used? So and then the IRB said you're exempt. How did you come up with these strategies? Why did you decide to use these strate-, these privacy protection strategies? And then how did you come up with what you would and wouldn't do in terms of protecting privacy?

BSR09 23:57
Okay, thank you. And I think that this process was started because of the IRB issues mentioned in previous papers. So some other researchers mentioned that and also, when we do the training to be a researcher in the university, they also mentioned the kind of, like, rules we have to follow. So both university training and also the previous papers, tell us we should do that.

Sara Mannheimer 24:31
Are those previous papers that you published or that you've read?

BSR09 24:35
Other researchers.

Sara Mannheimer 24:36
Yeah. And so did you discuss it within your research team? and say, "Okay, these are the ways that we're going to protect privacy," and did you end up publishing any parts of tweets in the paper? Sorry I'm asking a bunch of questions at the same time, did you consult with any about besides sort of reading different papers, and like the university training or whatever type of training you did, were there any other consultations that you did within your team or like with a librarian or anybody about, about privacy and how to support privacy with social media?

BSR09 25:20
Um, definitely, we discussed it with my advisor, and my advisor may discuss with librarians or other staff to talk about this issue. But actually, I do not have too much experience talking to other people. Yeah.

Sara Mannheimer 25:35
Yeah. And then did you? What did you end up publishing? Did you publish any quotes or examples of the tweets in your paper?

BSR09 25:47
Yes, we have some we included some content content on the tweets, but we never show like a phone numbers or address or any other information. In the paper we just show. For example, this house is damaged. We just chose this sentence. Yeah.

Sara Mannheimer 26:08
Okay, so just like snippets of the tweets? Yeah. Okay. Awesome. All right. Last question. Intellectual property. Can you tell me about a time during this research when you considered intellectual property concerns? So you talked a little bit about the Twitter Terms of Service? I feel that's one of their Twitter's ways of protecting their... the intellectual property of the company, I guess. Were there any other concerns that you had around intellectual property?

BSR09 26:52
Wondering how can we define intellectual property?

Sara Mannheimer 26:56
I guess it would be like, well, with your example, because you purchased the data, I feel the intellectual property concerns are actually pretty cut and dried. But thinking about, you know, like, people, when you say, when a person tweets something, that's if I make a tweet, that's my own intellectual property, you know, like what I said belongs to me. But then Twitter has a license to post that on their website. It's like they, because of the Terms of Service, I've given them the permission to post it on their website, and I've also given them permission to provide it to researchers. But one of the issues that I've seen as I've conducted this research is that Twitter users don't often fully read the Terms of Service and don't always completely understand what they're agreeing to. So I'm just wondering if....

BSR09 27:55
Oh sorry.

Sara Mannheimer 33:19
No worries. I was gonna write you an email, but I was like I think I'll just wait. Well, we're nearly done anyway, so oh, I see, I got an email from you this week, I guess I was just sort of talking through all the potential intellectual property, I guess mostly with social media research, it's their social media Terms of Service, and then the intellectual property of the Twitter users. So it's okay to say that you didn't consider it too.

BSR09 33:52
Um, I think the intellectual property might be the information they post related to their emotions or their reactions to the disasters. So we actually we have a paper to study the emotion of the users on social media to see how the emotion change can reflect the damages of the infrastructure systems.

Sara Mannheimer 34:23
Oh, I'll check that one out. Okay. All right. Great. Well, are there any other issues or challenges that arose during this research that I haven't asked you about? I think that's all I have on my interview guide. I have a last question about looking for other people to interview but you're one of my last interviews so

I don't need any more participants. Thank you so much. I really appreciate it. And good luck with it, the rest of your research and the rest of your postdoc.

BSR09 35:03
Thank you. Thank you. Good luck research as well.

Sara Mannheimer 35:07
Thank you so much. Talk to you later.

BSR09 35:10
Bye