BSR10_transcript_deidentified

SUMMARY KEYWORDS
data, tweets, people, twitter, fake news, collected, research, topic modeling, question, project, social media, bots, issue, develop, intellectual property, natural disaster

SPEAKERS
Sara Mannheimer, BSR10

Sara Mannheimer 00:00
Okay. So yeah I'm Sara Mannheimer. And that's my background I'm... the research, the main research question is how can we data curators best handle qualitative and big social research so that we can support responsible, ethical epistemologically sound, and legal share data sharing practices. And so the interview today is structured around six key issues which identified through a lit review. We have context, data quality data comparability, informed consent, privacy and intellectual property. And so I'll ask one question for each of those. And then I have an intro question and a wrap up question as well. So it should take us about one hour or less. So my first question for you is, introduce yourself and tell me about the type of research that you tend to do, and what type of data you produce.

BSR10 01:05
Sure, currently, I'm an [faculty member at a university]. And I [am in X department]. Okay. And I, I have a background in computer science. And while working on my PhD work, [there was a large-scale crisis in the U.S.]. So I was a graduate student, and my advisor at the time, worked on a project about archiving all these [posts about the crisis], like social media, usually Twitter, and along with the associated webpages that people share through Twitter, Twitter feed. So I became the GRA for that project, which lasted for many years. And I was a guardian to this in a world of social data collection, and analysis to some extent, and developing all these archives, by usually webpage archives, based on social media data, by working with the Internet Archive. So I've been in that field, like uh crisis informatics for a while. And since then, you know, even after I came to [my current position], I, I could work with other people from like, all different departments like environmental science, social, social science, sociology, and political science and computer science, to to do a [grant funded] project about this using social media data, in addition to traditional like census data to identify all this, like, community resilience, when all these natural disasters hit. So I've been in this field for a while using a kind of social media data.

Sara Mannheimer 03:45
That's so interesting. Can you so I guess, did you I had asked you at the beginning or when I first emailed, you have some time ago, to think of a recent time when you collected social media data or reuse data or prepared it for sharing? Did do you have an example in mind and we don't have to stick to it the whole time. But it helps to have a specific example as we talk through these questions.

BSR10 04:12
More... the most recent case, we'll be collecting Twitter data about [a recent national disaster]. [Additional details redacted]. And you know, whenever I experienced or see this type of natural disaster I'm interested in like, how people communicate and who are the major people in this like social network,

you know, sharing and trying to spread information and all this stuff. So I started collecting these Twitter feeds from [a certain date]. And [for about a month], and yeah, and then I am using this data like collected using keyword [company name redacted], which is an electrical company [in the state where the national disaster occurred], and also [the national disaster] hashtag, I could collected a couple hundred thousand tweets. So I started working on that data set to look at social networks and some topic modeling and stuff like that. Yeah.

Sara Mannheimer 06:04
Um, so when you were collecting the data, did you have a specific research question in mind? Or were you just since this is sort of your area? Do you just collect and then come up with the research question, as you go...?

BSR10 06:17
A research question? Um, what kind of a basic standard research quest-, question for me, in this crisis informatics will be like, I usually think about "what, when, where, and who" aspects of this disaster. And so, you know, this disaster had a before phase, a during, and an after phase, based on how much this, you know, [natural disaster] affected certain areas. And I, I'm interested in like what are the major, like, significant central people in this social network, and what they talked about and what kind of information they spread out. And, and their location information, too. So I will I plan to use, like, ArcGIS to plot their locations and see how that looks. And when will be like, three different phases? So, yeah, I usually look at those as a baseline.

Sara Mannheimer 07:36
Nice. Have you found that there, you can often find geographical information, like location information on the tweets?

BSR10 07:45
You know, it's really rare that people trans-, turn on that feature. So I'm trying to rely on their user profile data. Because some people, you know, add their location data, like I'm in [city, state], and stuff like that. It would be like an approximation, but you know, no choice.

Sara Mannheimer 08:13
Yeah. Okay, let's move to our first theme here -- issue, which is context. So just a quick sort of overview of what I mean by context. When we collect data from social media platforms, just like when we collect data from traditional spaces, the context is important. But the context of social media posts may be absent or difficult to understand for, you know, for example, that location information might not be there other contextual clues might not be available. And also, social media posts are short pieces of text or images or videos that are kind of taken from a larger context of personal or public life. And then when you're amassing data at a large scale, it can become even more complex. So can you tell me about a time if any, during your data collection for this project, when you considered the issue of maintaining or understanding the data's context?

BSR10 09:17
Context? That's a good question. Um, if we look at the data, like each individual, social media like tweets in a macroscopic way, then it's hard to understand sometimes because there will be a lot of junk posts, advertisement and fake news and fake news a big problem nowadays, and some people and bots will use a popular, popular Twitter like hashtags, just to advertise their product to so completely non-relevant content could be collected in the social media field. So I'm trying to look at in a little more macroscopic way to, you know, to do this, I usually apply like topic modeling, to see what kind of topics

comes up from this, the whole collection of this Twitter data, like, hundreds of thousands of Twitter data. And yeah, so also, I typically to understand some kind of large chunk of Twitter, I often use, what is the word.... a word cloud. Yeah, to quickly make the word cloud, and then see what kind of the main concepts and keywords people talk about?

Sara Mannheimer 10:05
Yeah. Yeah. Do you...

BSR10 10:32
....I understand the context.

Sara Mannheimer 11:03
Nice. Um, do you do any, like, filtering out of bots, or bad quality tweets? Do you have sort of automated methods that you can use to sort of clean the bad data out of there?

BSR10 11:19
There will be all-, always issue and I don't know for now, I didn't do much on that. That's one issue too. If I use a filtering, then I may remove some useful data, too.

Sara Mannheimer 11:42
Right. So you were just use-, with the topic modeling. I mean, obviously, the bots and the fake news accountants wouldn't be top-, using talking about the same topics as others on the hashtag. So maybe they kind of filtered themselves out automatically. Oh, yeah. In that way.

BSR10 12:01
If if I just run a topic modeling for the unfiltered, unfiltered tweets, I will have, like a couple topics. All, you know, composed of like how you URLs, https and all these, because that's part of the tweet. So I know how to remove like URLs. And, and for those I don't know, I try to look at these a social network to see whether there's, like really influential entity in the network. And whether it is more like human being or bots, or if something else some organization by doing a social network analysis.

Sara Mannheimer 12:58
Oh, so do you do all these strategies at once, like, do you do you do the word cloud? You do the social network analysis? And then also the topic modeling?

BSR10 13:08
Yeah, I try to do that.

Sara Mannheimer 13:10
Okay, yeah.

BSR10 13:11
Because, uh, you know, this natural disaster is like, has a multi-faceted data set. So I want to see a little better picture from different angles. People involved in this and their topics. And their like, locations. Yeah.

Sara Mannheimer 13:36
Yeah. Okay. location and network. Awesome. One more question, did you when thinking about these, like ways that you've sort of worked out to better understand the context of the data? Did you consult

with anyone or look to other research projects? Or.... I don't know, look to any policies or guidelines, what's been your strategy to develop these methods that you use to sort of better understand the context of the data?

BSR10 14:09
I think, yeah, over time, since my graduate student years, I could I could work well, all my projects were multidisciplinary. So I had many chances to learn from sociologists and environmental scientists, geologists and people from many different fields. So over time, I kind of developed my current strategy and a set of tools to look at this social media data.

Sara Mannheimer 14:49
Great. Okay. We talked a little bit about data quality with the bots but can you tell me about a time, if there were any, during this research project when you considered the issue of data quality? So it could be like missing data or bias or bots, or the quality of the method to answer your questions.

BSR10 15:16
For Twitter data unless it's purchased it's, as you may know, you know, it's a really small kind of random, small percentage of kind of random sample, if we collect them using a Twitter API. So well, at this point, I kind of use all of them, even though, you know, there will be some posts by bots, I kind of include them as part of the data set and see whether but could be, you know, kind of influential central entity in the social network. And in most cases, if, if it's not become so popular in the network, then it'll be kind of, you know, not distinguishable from other central people central figure. And if they are identified as one of the central figures in the network, then I want to look at it in more closely and see how they do and stuff like that.

Sara Mannheimer 16:33
Right. Like there might be a bot that's tweeting out like the most up to date [information relevant to the disaster] or something like that. And that does, that wouldn't, that would still be something you were interested in. Okay. Any other data quality issues that you've encountered during your work?

BSR10 16:50
Twitter is full of junk, junk messages, too. And, well, recently, I specifically looked at the fake news. Yeah, this information and misinformation occurring in Twitter. So I had that as a separate study. And, yeah, based on these snopes.com, the list of fake news, during a certain disaster. And specifically collected a list of fake news that were spreading during this [specific disaster]. And then I look at those social media cascade stream of these tweets connected with each other specifically for that disaster. And yeah, so I look to

Sara Mannheimer 17:57
Did you find-...

BSR10 17:58
Oh, sorry.

Sara Mannheimer 17:59
Did you find that those tweets influenced other tweets? As you were doing that research? The disinformation?

BSR10 18:09

Uh, tweets influence other people?

Sara Mannheimer 18:14
Yeah, like, did the tweets with disinformation or misinformation, get spread? And then did it influence? I guess, I don't know. What what were your results, was all I'm asking.

BSR10 18:24
Oh, yeah. Results. Um, you... I could see many political political tweets. Yeah, from like, around the [year when the disaster occurred]. There was like, a lot of, like, stuff by, like, left and right, by conservatives and more democratic, Republican and all these things, so I could see some fake news generated by far right and far kind of far left also. And, you know, a lot, a lot of stuff about [details redacted]. These uh, activists saying that they were some people were blocking right process. Some recovery activities and stuff like that.

Sara Mannheimer 19:45
Hmm. Interesting.

BSR10 19:48
Yeah. And usually those tweets become really popular in the first one or two days. So if you make a graph it just skyrockets, like in the first couple of days. And then after two or three days later, I could see that people who are skeptical about the news may post this on Twitter, like, referencing a fact-checking website.... Oh, actually what you're talking about is fake news. And they were trying to correct those invasive spread of misinformation. So I thought that's an interesting kind of self regulatory activities behaviors in a social network.

Sara Mannheimer 20:43
Yeah.

BSR10 20:44
So yeah, that's one interesting thing from the fake news and social media.

Sara Mannheimer 20:49
Wow. So interesting to be able to study this. Well, I feel like this is somewhat related, or the next issue is data comparability. So in the [two examples you've discussed so far], did you compare or combine multiple data sets? So you said you used Snopes information during the [second] study? Can you tell me more about like, if you encountered how, what challenges you encountered when you're been trying to combine multiple datasets?

BSR10 21:27
Combine multiple data sets...

Sara Mannheimer 21:29
Or use them in, in conjunction like you would talked about using census data with the Twitter data and stuff like?

BSR10 21:38
Well, combining well in social Twitter data, usually we use, you know, multiple keywords and multiple hashtags generated for a specific natural disaster. So for example, I will use a hashtag [hashtag redacted], either keyword [keywords redacted], some other keywords too. And then once we develop all

these keywords, I the last thing, the next thing will be to kind of merge them together. Because people use multiple hashtags and keywords at the same time. So merging was the one issue.

Sara Mannheimer 22:25
Was it difficult to merge because of the size of the data? Or what were the difficulties there?

BSR10 22:31
Well, yeah, the size of the data, somehow we need to put it in the data database. And by using just probably spent, well, if we use a database, technically, it's not a too much of a big deal, because each tweet will have its own tweet ID. So we can, you know, color code it so that the database can have only unique tweet IDs. So but you know, how if you work with people, different institutions, for example, for either of these collections, I was trying to share them with [a different university], some of the people [at that university], because they, they could work with the Internet Archive, and I was trying to help them develop web page collections. So in the case here, it's big data, somehow I needed to put it somewhere like in the cloud or in my own server, so that people can download that. At least, you know, Twitter IDs. So yeah, I'd say size of the data matters sometimes.

Sara Mannheimer 23:59
Yeah. How about when you're using like census data? Or Snopes data? In addition to the Twitter data? Do you encounter any challenges there?

BSR10 24:16
I don't personally use a census data.

Sara Mannheimer 24:20
Okay.

BSR10 24:21
Yeah, but my yeah, my team members use the census data. Yeah. And for using snopes.com. For the [natural disaster], we have like millions of Twitter data collected. And then I wanted to find the specific tweet, that reference of specific fake news. So I have to search through all these Twitter data in in my database. Then to do that snopes.com will have its own kind of title for that fake news event. Okay. But if you use that title as is, you will have a kind of... How would I say? you will only collect tweets that reference snopes.com exactly as the title says. So I had to develop some strategy to use some synonyms of some certain keywords of these titles of the fake news and stuff like that.

Sara Mannheimer 25:45
Yeah, nice. So, um, was this, the fake news or the disinformation study you did with [the disaster]? That was on data... Because the article that I had found you through was this one where you're looking at police activity during [the disaster], police activity on Twitter? So did you use the same data set for those two? Is it like a big data set [about the disaster] that you've collected, and then that you're using for different purposes?

BSR10 26:12
That data was smaller for that data, I collected it myself using a tool. And for the fake news data, we purchased the [disaster] data, with our funding, from Twitter. So that was a more more complete kind of data set. Okay, for the police study, I use the data set that I collected using the Twitter API.

Sara Mannheimer 26:46

Nice, okay. Yeah. Cool. All right, let's move to informed consent. Can you tell me about a time if any, during this research when you considered informed consent, participants so to speak like the Twitter users?

BSR10 27:07
Well, not specifically for me, but a one of my team members, project team members, they use informed consent when they want to... when they were trying to do surveys for the people. And they collected the contact info, these people's name or contact information from the Twitter data set. So it started from collecting the Twitter data set. And they searched through some of interesting figures in Twitter, Twitter, including not only individuals, but also organizations, and like NGOs, [and other groups active in the area where the disaster occurred]. And then when they wanted to conduct surveys, they went through all these informed consent and stuff like that. Yeah. Yeah.

Sara Mannheimer 28:10
And contacted them? Okay. Yeah. Tell. So did you go through IRB or any processes when you were doing this Twitter research? Or were you just thinking of it as existing data that didn't require human subjects treatment?

BSR10 28:25
Yeah, Twitter data. I didn't go through IRB.

Sara Mannheimer 28:29
Yeah. And did you? Like, did you feel that the participant or the Twitter users whose tweets you collected...

BSR10 28:42
Yeah.

Sara Mannheimer 28:42
Do you feel like they would expect to be used in or in, in research like this? Or what did you feel their expectations were as far as being part of research or not?

BSR10 28:57
It's a kind of a debatable topic.

Sara Mannheimer 29:00
And that's why I'm asking about it.

BSR10 29:04
Well, if I go back to another study, we, with another, my colleague at [my university], we collected a tweet posted by the libraries in [disaster areas] and then wanted to look at how they communicated during a certain like [disasters]. And there were some just library patrons communicating with these libraries. So in the case, when I publish things, I try not to focus on these individuals.

Sara Mannheimer 29:49
Yeah.

BSR10 29:50

And try not to talk about what they said and stuff like that at all. And my focus will be more on like, you know, these more public entities, like institutions, federal entities, like public libraries, and FEMA and stuff like that. So even though I use Twitter, and maybe some of these, their tweets will be contributing to my topic modeling study. But I try not to talk about, like individual tweets, exposing their private information like that.

Sara Mannheimer 30:31
Okay. So you don't do any, like quoting of tweets, in your articles?

BSR10 30:39
Quoting?

Sara Mannheimer 30:42
Like, would you say, you know, this person said, "[disaster is occurring], and my electricity is out? Or something like that, you know,

BSR10 30:49
I try not to, at least don't provide the, you know, like their Twitter IDs, or something like that. I try not to, unless it's really...

Sara Mannheimer 31:01
Tell me more about like that thought process? How did you come to that decision to sort of focus more on the institutions rather than individuals?

BSR10 31:11
Yeah, it's a well, now, you know, Twitter provides the Twitter data for free for research purposes now. And so that means, you know, as long as we abide by their terms of service, and we can use the data set in a research and but well, I've attended the training for data curation provided by [a university] a while ago. So including those experiences, I learned about, you know, this, protecting privacy is, you know, really important matter in your kind of social social data set, or any, like medical data set to specifically. So by also talking, discussing with my colleagues, I could sort of develop that. Not only myself, but you know, it's kind of a team team effort. I uh, I think we, we should better do this way or that way.

Sara Mannheimer 32:32
Yeah.

BSR10 32:33
Over time.

Sara Mannheimer 32:34
Yeah. Okay, that's great. That kind of led into my next question is about privacy. So oh, maybe we've covered this. But do you have any other thoughts about issues of privacy, like protecting the data during research? Or do you ever make your the Twitter datasets you have available to other researchers? Or have you ever thought about publishing these datasets?

BSR10 33:02
For Twitter, we, in my project team, we share we purchased maybe total three data sets from Twitter about [three different natural disasters], and then since we talked about, we will share some of our data

set. When we submitted our proposal to [a funding agency], uh, we share the tweet IDs but it's like 1% or 5% of tweet IDs and as a sample Twitter data so that people are interested in this could hydrate the full Twitter data.

Sara Mannheimer 33:54
So you went... you got a grant from [a funder] to pay for the three data sets that you bought and so then was it because you weren't allowed by Twitter to publish the whole thing that you'd purchase the tweet IDs you just... they is that like from the terms of service they like only allow you to publish a certain percentage of tweet IDs?

BSR10 34:18
I don't think there's a restriction on that.

Sara Mannheimer 34:22
Okay.

BSR10 34:23
But that was our decision Yeah, if yeah, I believe we just to share the some sample Tweets and since one data set probably [certain disaster] was like 16 million tweets, so 1% will be fairly many tweets. So yeah. Yeah. So if we just to share the, well, I'm more like, open sharing person. Yeah. But then after talking with with my other group members, we finally decided, let's share some small percentage of the actual data set that we use.

Sara Mannheimer 35:10
And where did you share it? Did you share it in a repository? Or as a kind of...?

BSR10 35:16
Oh, yeah. Through our project homepage. Yeah. I don't know if you're interested in that.

Sara Mannheimer 35:29
Sure, if you have it handy yet, you could put it in the chat.

BSR10 35:32
Yeah, let me show. Yeah, that's a specifically data sets page. And we share the some of the web pages we collected by collaborating with [another university] and also, we have [tweets from other natural disasters]. Yeah, these are like samples.

Sara Mannheimer 36:02
Nice. Perfect. Oh, that's cool. All right. Um, anything else you want to add about privacy? So like you said, your main approach to privacy was to try to talk about the tweets in the aggregate, but never speak of individual people or use individuals, quotes from individuals tweets and focus more on the institution. Any other like, thoughts that you heard about privacy as you've conducted these research projects?

BSR10 36:35
And this is actually a big question to me too, because sometimes, for the fake news study, there were certain individuals who spread, who are spreading this false news on a big scale. Thousands of people were retweeting. So, at that time, I kind of said, oh, this person, I, I kind of talked about this person, saying, oh, this was a, this person is located in [country], and you know, based on that person's the

public user profile data bio, that they typed in by themself. So I kind of, you know, I'm not on like, left or right. So I kind of described the situation, or this person, for some reason, posted this false news. And 1000s of people were retweeted in the beginning, but then I could see another group of people were sharing the debunking article from snopes.com after a couple of days later, and also some visited this person's Twitter page and left a message "Hey, you are spreading fake news." So I thought the whole event was interesting, like a self regulatory behaviors online. So in that time, I, like, I used this person's Twitter username, oh, [Name] blah, blah, and yeah. But if there is a I don't know, some other way that I don't have to do this. I will follow that.

Sara Mannheimer 38:31
Yeah. Do you, like.... Are you aware of any like, have you looked at any guidelines, like when you decided, Okay, I'm going to publish this person's Twitter username. I'm like looking to hear more about people's thought processes, because there's just such a variety of ways that people think about this stuff. You notice there's no like real good... There's no IRB for social media data, you know?

BSR10 38:55
Yeah.

Sara Mannheimer 38:55
So how did you... how did you work through those ideas? Did you talk to people? Were there other similar studies that might have also used individuals ID Twitter usernames, or what was your? How did you come to that conclusion?

BSR10 39:11
I usually I work with other people when I write a paper.

Sara Mannheimer 39:18
Yeah.

BSR10 39:19
So I talk to my collaborators first, which will be a better way to do this. And I also talked with my project team members about all these social media data. And always the dilemma... kind of a dilemma.

Sara Mannheimer 39:45
Yeah, I am hoping through this project, we'll be able to develop some sort of more standardized practices that curators and librarians can suggest. So this is really helpful. Thanks. All right, our last question about intellectual property. Can you tell me about a time, if any, during your recent research when you considered intellectual property concerns? So like, for example, looking at the social media Terms of Service or thinking about intellectual property for publications or news outlets or or people in the Twitter data set?

BSR10 40:30
That I'm actually I haven't thought too much about this intellectual property aspect of social media data.

Sara Mannheimer 40:40
Yeah. Okay. That's a good, that's a good answer. But you have you did read the Twitter Terms of Service, you were saying, right, thinking about?

BSR10 40:55

Yeah. And one thing is that they change often, they change often. So it's sometimes hard to keep up with the update. And when I talk to colleagues in other institutions, because I work with, you know, writing all these proposals, I work with people at [various universities]. So sometimes different people talk a little different stuff. So like, what, I don't know, some little bit outdated things. And, I don't know, for social-- sociologists may have a little better sense of this. And if I talk to engineers, and computer scientists, they were more on like a technical aspect of data collection and analysis. So less time will be given to looking at all these terms of service and stuff like that. Sometimes by working with the people who, who knows more about this intellectual property and privacy issues in detail, that helps as the project team as a whole, yeah.

Sara Mannheimer 42:29
When you're putting together a project team, do you try and include, like a social scientist on it for that reason? Or do you mostly consult with them?

BSR10 42:38
Um, project team? Well, I happen to join a project team for my prior [grant funded] project, through my [university]. And it's, I don't think I specifically selected social scientists for that reason. And I bought, I used to work with social scientists who were interested in like crisis informatics and community resilience and those topics, so.

Sara Mannheimer 43:23
Yeah.

BSR10 43:25
But I don't specifically pick people from sociology.

Sara Mannheimer 43:31
Yeah. Interesting. Okay. Any other issues or challenges that arose during your example, your research that I haven't asked you about?

BSR10 43:45
Other issues? Well, still, nowadays, some journals will ask you to submit the data set that you use the for the paper. So to some extent, I will kind of curate, like a Twitter data by myself, or only share Twitter IDs sometimes. But then sharing is a still a little issue, sharing it properly. Yeah, because, uh, I have my own server computer and I, I use a server or my laptop. So for a while, I'll store this data in my server. But then, I don't know it's hard to do, like long time sharing. I can maybe use a, you know, good public data repository like Zenodo. But it it requires my time and effort which is uh, natural.

Sara Mannheimer 45:02
Yeah.

BSR10 45:03
Sometimes not easy to spend time on that.

Sara Mannheimer 45:07
Yeah. Yeah, I'm experiencing that as I'm thinking about, oh, I want to publish these transcripts, I want to share my data and practice what I preach. But it is it's like you want to get your paper out? And if it's not a requirement, it can be tough. Yeah. Have you encountered any issues when you share your data on

the website? Like, do people contact you? Or what's been your experience with those subsets of data that you've shared?

BSR10 45:37
Not much issues actually. There were people who downloaded our data set, but then not much issues.

Sara Mannheimer 45:50
And how long do you tend to keep the your data? Do you keep it indefinitely? Like on your own personal servers?

BSR10 45:59
Um, indefinitely? Yeah, at least three, four or five years?

Sara Mannheimer 46:08
Yeah.

BSR10 46:10
But at some point, I will just delete all the data I don't use and, yeah, I usually backup my data. But then after some years, I just clean out.

Sara Mannheimer 46:25
Yeah, yeah. Okay. Well, I think that's it. You're my final interview.

BSR10 46:33
Oh final!

Sara Mannheimer 46:35
Yeah, thanks. I should be finishing by April. That's my plan. Analyze the data. Yeah. Yeah.

BSR10 46:47
Good luck.

Sara Mannheimer 46:48
Thank you so much. I really appreciate you taking the time to talk to me. It's been really interesting to hear your experiences.