BSR01_transcript_deidentified

SUMMARY KEYWORDS
data, people, platform, project, reddit, research, questions, big, paper, api, user, context, moderators, debates, posts, review, repository, gender, community, findings

SPEAKERS
BSR01, Sara Mannheimer

BSR01 00:00
I haven't looked at the questions you sent. Yeah, but this is an active project. So I don't think I should have a problem with any of them. Okay, great.

Sara Mannheimer 00:13
And do you have? Do you want to follow along? I can resend the email that had the attachments, if you like. That may help you.

BSR01 00:23
I have the email. Yes. Let me just open it up.

Sara Mannheimer 00:26
And you don't have to I can also just speak, but it's sometimes nice to have it to read as well, since some of the questions are more complex.

BSR01 00:35
The big social interview dot, dot PDF.

Sara Mannheimer 00:40
Yep, that's it. Yep. So hi, my name is Sara. I'm a librarian at Montana State University. And so the reason I'm doing this research is trying to figure out you know, big social data is a newer practice. And we have a lot of questions about data curation surrounding big social data. So the idea behind this project is that I'm trying to connect the communities of qualitative researchers and big social data researchers, and the data curation practices that are done in each community to try to connect the two and hopefully learn some things, maybe about scaling up for qualitative research, and maybe about better data management practices in big social research, and more ethical and epistemologically sound research. So, so yeah, let's get started. Can you tell me a bit about the type of research you generally do and what kind of data you produce?

BSR01 01:43
Yes, um, so my research is, there are two types. I work on methodological research developing methods to analyze conversations, to do things like detect fraudulent reviews, fraudulent transactions, things like that. I also work on understanding human behavior in conversations using big online conversational data sets. And one of the projects I've been working on for a while is understanding conversations on this platform called [Online Debate Platform]. And I'm specifically looking at what is the impact of status indicators on persuasion on this platform. And various status indicators give you

enhanced persuasive power, just by virtue of themselves. And also to what extent do status indicators motivate people to participate? Because it's a voluntary platform. From the platform's perspective, you want people to participate. But you also don't want them to have any special advantage. So I'm trying to tease apart: how do these two things go together? And what platforms can do about it.

Sara Mannheimer 03:06
Yeah, the paper that I read was really interesting. I didn't understand all of the methods, so I'm glad you agreed to talk. So you've identified this specific example. And you're going to use the one with the [Online Debate Platform] and understanding behaviors there? Okay. So can you describe your data collection method from that platform?

BSR01 03:34
Yeah. So there are two types of data I'm collecting from this platform, the first is just archival or historical conversation data. So this is like, it's data going back to [the 2010s], when the platform first started, to today. And it's like around millions of conversations. The way I collect this data is, there are three sources actually, the first is using this archiving platform called Push Shift. Or what Push Shift does is it archives everything on Reddit

Sara Mannheimer 04:14
Is it push shift? S H I F T?

BSR01 04:18
Yeah, Push Shift. So they archive everything on Reddit using... they figured out a trick, which is basically every Reddit post or comment ID is just basically an integer. It's a number incremented, one by one. So what they do is they get all of them and store them somewhere. Push Shift is nice, but it's it has a lot of holes, there are no guarantees that your data is complete, which I realized because I was reconstructing these conversation trees from the data and then I realized that some comments are missing. So a second data source that I use is the official Reddit API that's provided by Reddit. And it's like, you can give it a comment ID or a post ID, and it gives you back that specific comment. But there are a bunch of limitations, you can only access the API once per second. If you're looking for the most recent comments by a user, you'll only get 1000 at most. So that's the second source. The third source is something called the Push Shift beta API. It's like the Push Shift API. But it's it's just, I don't know what exactly the difference is, it's just in a different location, and sometimes Has more, or some of the missing data is available in this source. So those are three data sources. And I collect them all using some kind of Python script that connects to the API, fetches the data and stores it on on my machine.

Sara Mannheimer 06:04
Yeah, we'll talk about missing data later on. And actually, why don't we just talk about it now, since we were you've talked about your like, strategies for pulling in missing data, we're also going to talk about data quality. So how you account for data quality? An example of when you considered that issue? So you've talked about getting missing data from the API's. And you've also talked about your methods being used to detect fraudulent posts or bots. But what other examples do you have about data quality issues in your research?

BSR01 06:46
Um, so one example is missing data. Another example is, I work a lot with text data. So sometimes parsing the text data that's online, into something that's human readable. That could cause issues, especially when people use complex—these Unicode characters, basically, from different languages.

Sara Mannheimer 07:14
Yeah.

BSR01 07:14
When you try to read that, in many programming languages, they just, they're just removed, like those characters are just removed, and things don't make any sense. So that's a big issue with text data.

Sara Mannheimer 07:29
What do you do to account for that? Is there a way for that you've been able to correct those mistakes?

BSR01 07:36
In my specific project, I haven't needed to yet. But there is a way. Uh, you'd have to treat the text as this format called Unicode. It's a, it's a, it's a character set that can accommodate a bunch of languages and special characters. So I'd have to treat the text as Unicode and that should fix things. Okay. Yeah, it's a bit more effort from my side.

Sara Mannheimer 08:04
And do you share the data that, I think this question is back where we were, but do you share the data that you pull down from the Reddit API? With your paper or in other places?

BSR01 08:18
Yes, yes, I usually do, I usually share the data once the paper's published, or because then I'm like, I'm sure that the data is in its final form, and it's clean. And it also replicates the results in my paper. I have some.... So if you just, if you go to my website, the datasets for all my previous papers have been released.

Sara Mannheimer 08:43
Okay. And do you... Where do you release them, just on GitHub?

BSR01 08:49
I yes, I release them on GitHub. Sometimes. So GitHub is not good for fairly big data sets, which is what my data—it might change—my data is right now. So I am I'm trying to find a better place to share that data set. I might just share it on my website as a raw download. I'd have to find a good a better platform.

Sara Mannheimer 09:16
Have you heard of Zenodo?

BSR01 09:20
No, no.

Sara Mannheimer 09:20
I'm not sure it's through CERN, which is a Particle Physics Laboratory. Z E N O D O. It's a data repository. I think they can accommodate big files, but I'm not sure. Maybe it's.... Yeah, I think you'd have to check with them. You might have to like have a special, some special agreement. But that's a good sort of just general purpose repository, if you're looking for something.

BSR01 09:38

Yeah, there's that. Thanks for that. I've also heard of Harvard Dataverse. I haven't used these platforms. The issue with uploading stuff on these platforms is they don't show up on search results most of the times. So people won't stumble onto your data sets the way they would on Github. Kaggle, Kaggle is another place where I could upload it.

Sara Mannheimer 10:07
That's great. That's a good idea. Yeah, from a library perspective, we're like thinking about preservation as a data. And so the data repository has a DOI. So it would potentially show up. If you're searching like Google, it wouldn't show up in Google Scholar. But have you seen Google data set search? That's a sort of.... I guess it just came out of beta. So it would show up there, but Zenodo and other like data specific repositories that are about preservation, so and you can actually connect Zenodo with GitHub to get a DUI for your repo. But anyway, this is the beside the point.

BSR01 10:53
This is useful.

Sara Mannheimer 10:56
Okay, and maybe [your institution], if you ask the librarians there, they might be able to help you with it, too. They might have a data repository locally.

BSR01 11:09
Yes. Yeah, I should, I should check with them for sure.

Sara Mannheimer 11:13
And so is your project grant funded? Were you required to share the data? Or is this just something that you're interested in yourself?

BSR01 11:20
No, these are not grant funded. Yeah. These are just funded by my PhD stipend.

Sara Mannheimer 11:35
Okay, let's move on to context. So I have this quote here. "When we collect data from social media platforms, context matters. But the context of social media might be absent or difficult to understand if you're taking short pieces of text out of the broader context of who the person is, what their personal and private life or in public life are. And then when the data are masked at a large scale that can be even more pronounced." So when I talk about context, that's basically what I'm thinking about. So can you tell me about a time, if any, during your process, when you considered the issue of how you might maintain and understand your data as context, both for you doing the research and for, potentially for users when you share the data?

BSR01 12:21
Yes, yeah, context was... So when I submitted this, the first project from this data set to a journal, it was the conversations between users on [Online Debate Platform], but I was trying to answer a managerial question or a question or a question about human behavior in general. So I needed to qualify my findings with what is the population I'm looking at. I would have liked to know, things like what is the age, age group that people are in? What topics are they interested in debating? Why do they participate on this platform? Things that if I, if I could do a survey of these users, I would ask them a lot of questions. But I can't. And all the users are anonymous in the sense that there is no identifying information associated with their usernames. So I actually don't, I don't have stated context from the

platform. The way I worked around this was to look at overall Reddit demographics. And assume that people on this part of Reddit on this subreddit would be similar to the overall demographics. That's, that's a pretty strong assumption. It was just kind of the best I had.

Sara Mannheimer 13:47
And did you talk about, you know, when you publish your data, or you publish your paper, do you talk about that in the paper as like a limitation? Or do you have a discussion of it?

BSR01 13:58
It's something it's not there right now. It's a, it's a comment that a reviewer brought up. So, we'll definitely... Yeah, it's going to be like a limitation or something that's going to be in the conclusion talking about the external validity of, of our findings.

Sara Mannheimer 14:15
Okay. And then did you, besides the review, or after the reviewer brought it up, did you like consult with anyone else or look at other research projects? How did you come to your thoughts about how to present context and your data?

BSR01 14:31
Yeah, I did look at other papers that use [Online Debate Platform] data, or that use Reddit data. I guess specifically in the journal that I was trying to publish in. Because in, in, so in CS, in computer science communities, it's not that important, because everyone just assumes that your findings are limited to some specific context. But in management journals, it's a lot more important. So you see people talking about the context.

Sara Mannheimer 15:07
Okay. That's interesting. That's part of my, what I'm wondering is like how different communities, different disciplines relate to these questions. So that's super helpful.

BSR01 15:21
Yeah, so I publish in both, in computer science and management. So yeah, I can I can talk about it. Quite a few differences.

Sara Mannheimer 15:30
Okay, cool. Yeah, I have... So I meant to tell you, there are six different issues that we're going to go through context, data quality, data comparability, consent, confidentiality, and intellectual property. And so you'll have more opportunities there to sort of explore it. These are issues that I identified in both qualitative data and big social data. Okay, let's, and there are eight questions total when we're on number three. So, during your example, during your project, did you compare or combine multiple big data sets? Or have you considered comparability or interoperability of your Reddit data set?

BSR01 16:12
Um, I didn't. So far, I've not had to. So I collected the same data from multiple sources. I didn't have to combine it with other datasets, there was another project, where I was looking at biases in the text of peer review, where I collected data from an online conference review portal and I merged it with data from ArXiv, which is a paper archive repository. That was the only time where I needed to do this merging, but not in this present project.

Sara Mannheimer 16:54

Okay. Well, we can I think we can discuss that time, we can just switch it out. What strategies did you use to combine the datasets? Did you have to do any sort of adjustment of the data format? Or metadata?

BSR01 17:08
Yeah, it was pretty difficult. So I had papers on that were submitted to this conference called [Conference Name], that I got from this [online conference review portal]. I needed to match those papers to papers on ArXiv. Um, so the first challenge is, how do I even find these, I could look for the paper title. Sometimes people upload it with different titles, to avoid reviewers finding it and then getting to know who the authors are because the conference was double blind, in, after 2018. Now, I'd have to do things like do a search for the abstract. Even the abstract could, would not match that some conferences have a policy that you can't use the same abstract in your public key visible version of the paper. So you have to do something like a fuzzy text match. The good thing about this data set was it was small, so I could manually inspect every single match to make sure that it's right. So I could check for false positive matches, but not for false negative match. And so if I did not find a match, I didn't actually go and search for it manually.

Sara Mannheimer 18:25
Okay. Yeah, that's, this is a tricky one, especially the bigger the data gets, it's harder. I feel like there's a lot of potential there. But without common data standards that everyone's using, it can be really challenging.

BSR01 18:44
For bigger data sets, I usually plot simple statistics, like for [Online Debate Platform], I just plotted the number of comments and the number of posts every week, then that's what that's when I noticed that in February, there were like, a 10th of the amount of normal amount of posts, and that's when I knew that there was missing data.

Sara Mannheimer 19:04
Okay, nice. Great. Okay. Let's move on to question four about informed consent. Can you tell me about a time during your project that you considered informed consent?

BSR01 19:21
Yes, yes. Yeah. Yeah. So the [Online Debate Platform], I guess I didn't talk about it. I said the the first part of machine learning project was using historical data. The second part is a using a randomized experiment. So what I'm doing right now, what I just did, three months back was, I spoke to the moderators of the platform, who gave me permission to hide the reputation of anyone on the platform. And that way, that way I could do an experiment, I could randomly select 50% of the users hide their reputation and then observe their participation and their success, what successful persuasion rate over the course of three months. And then I can measure if there were any differences. And if there were differences, I could, I could attribute them to hiding reputation because their reputation was hidden randomly. That's what I did. I had to.... so before I started this experiment, I did have to get approval from the [Institution] IRB. And the IRB specifically asked me a lot of questions about consent. So they were interested in firstly, are the people on the platform going to know that you're hiding their reputation? And for me, it would be bad if they knew because that would change their behavior. So I didn't want to explicitly tell them. So I had to justify that this was not, not having informed consent while doing the experiment was not causing a lot of harm.

Sara Mannheimer 21:07

Okay.

BSR01 21:09
So that's, since this is a voluntary debating platform, and people are really participating, to get points and get better debating. I argued that there is there is no real harm in this, in this setting. So that was fine. That that worked for the IRB, though people did...

Sara Mannheimer 21:38
Yeah, go ahead.

BSR01 21:40
So people did, once I hid people's reputation, they did eventually find out. Because you kind of noticed, especially if you're an active participant. But yeah, and...

Sara Mannheimer 21:52
Did they tell the moderators that they hadn't seen that? How did you know that...?

BSR01 21:56
One of them did, one of them asked the moderators what happened, um, but I also saw that the people who I hid the reputation for—they reduced their participation, and the people who I did not try to hide the reputation for did not change.

Sara Mannheimer 22:14
So so it did affect their, it affected their standing in the community, maybe like maybe people weren't seeing their posts, or....

BSR01 22:22
It also affected, so it had a dual effect. It affected the number of debates they participated in, in those three months. So the people whose reputation was hidden, participated in on average, to fewer debates than the people in the control group. It also once you control for the number of debates, it also affected their success rate. So how many successful debates they had given a certain number of debates, they participated. So it had a dual effect on both of them.

Sara Mannheimer 23:00
So did you discuss like, Do you have plans? Or did you do like, what's your relationship with the moderator? Is, I'm wondering, like, if they could give you information about who they think their people are, what kind of consent they expect, or you know, what these what the research might, how it might affect them? You know?

BSR01 23:18
Yes, yeah, I'm so the, I created a Discord channel, to be in touch with one moderator who sort of my point of contact with the rest of the moderation team. I think they have around eight or 10 moderators. And I keep sharing my preliminary findings with this moderator and are getting feedback, talking about how this could help the platform. They seem pretty interested in the findings. And the the plan is once I write this up, they share it with the rest of the community.

Sara Mannheimer 23:53
Okay. Is there... So did you review the Reddit Terms of Service?

BSR01 24:01
Yes, yeah, yes, yes. Yeah, I reviewed them specifically for the case of collecting data. So they have certain terms, to use their API, things like that. I did not review them before doing the experiment. And that is because I had read quite a few papers that did their experiments. And I just assumed it was fine. I didn't get I didn't actually look.

Sara Mannheimer 24:33
Yeah, I'm not sure what their terms of service say either. Like I think that Facebook does specifically say when you sign up for the platform, you might get research done on you or with your with your data. Okay, cool. All right on to privacy and confidentiality. Can you tell me about a time if any during your research process when you considered issues of privacy, protecting the data during your research or protecting the people on the platform.

BSR01 25:03
Yeah, so a good example of that would be my previous project where I was looking at biases in the language of peer review. I was looking at gender bias, what are the biases looking with gender bias. For that we had an external annotator annotate gender for the authors on the platform of these papers. And the protocol we followed was semi automatic. So we use the US Social Security data to infer genders with some confidence probability. For example, if there was a certain name, let's say Jack, that was reported as male 99% of the time, we would tag it as male. And same thing if it was reported as female and 99% time we would tag it as female or non male rather. But if it was less confident, we had the human annotator, search for the name search for their Google profile and make a guess as to what this person's gender is. So we were, so we were trying to mimic the reviewers perception of the author's agenda. And because bias, the bias would be driven based on that perception, and not the self reported gender of the author. We did not release these gender annotations, because we didn't, we thought it would cause issues if we mislabeled an author's gender. And these authors are a part of the community that we publish. Yeah. So we opted to not release identities. And we actually at the end of the paper, we actually have an ethics statement where we talk about why we're not doing that. But the data, the data itself is public.

Sara Mannheimer 27:08
And I guess sometimes with social media, there are like considerations of move it, like moving the data out, out of its context, in Reddit, you know, and you're letting it be downloaded. I think in your situation, since all of the users were anonymous, that changes things a little.

BSR01 27:33
So they were anonymous in the sense that I couldn't tie it to their real identity. But some people use this same username on all these online platforms. And that is that that is a concern. So let's say there's a user in my data who said something they regret, and they delete their post or their comment. It's still going to be there in my data set.

Sara Mannheimer 27:57
Right.

BSR01 28:00
And that is, that is an issue that I have thought about, but I haven't figured out how to deal with it yet.

Sara Mannheimer 28:09
Yeah.

BSR01 28:12
I know that Push Shift has on the, Push Shift has a subReddit, where people can submit data deletion requests.

Sara Mannheimer 28:22
Okay. Interesting. So yeah, one of my questions is, did you feel that you had to make any compromises to privacy in order to conduct your research? Like, is this a major concern to you? Or do you feel it's small enough that it's not compromising user privacy?

BSR01 28:43
So it's not a compromise to my research. It's a compromise to replicating my research. So if, if I don't release the data, it will be private. But then no one can replicate my results. It's going to be really hard because you need to collect all this data again. Yeah. So that's the trade off that....

Sara Mannheimer 29:13
Yeah, like weighing the benefits of people being able to replicate and reuse your, replicate your research, and reuse your data, and then the potential privacy implications for the user. Cool. Great, great questions to be asking. All right. This is our last major question about intellectual property. Was there a time during the process of this project where you thought about intellectual property concerns of the participant or of the platform for anyone else?

BSR01 29:53
Um, so the, these two projects which I just spoke about peer review and [Online Debate Platform] didn't have any intellectual, I didn't have I didn't think about intellectual property concerns.

Sara Mannheimer 30:09
That's a good answer too. And then, are there any additional issues or challenges that arose during your example that I didn't ask you about or something you want to talk more about?

BSR01 30:28
So I think one, if one thing about missing data is the data is missing, because it was just wasn't archived. There could also be data that's missing because the authors deleted themselves from the platform. Okay. And that usually doesn't show up as missing data, it shows up as kind of text that is not associated with any author. So on Reddit, it shows up with the username of deleted.

Sara Mannheimer 31:08
Oh, okay, but the comment is still there?

BSR01 31:11
The comment is still there. And now I needed to think about what do I do with these comments? Do I add them in and treat them like they were made by a known, an unknown user? A unique unknown user? But then what if like, 100 of them are made by the same user, that would change my finding. So what I did in my project was to just drop all deleted comments. And also show that the results don't change much if you include them. Okay.

Sara Mannheimer 31:44
That's really interesting. And all of this, and you'll, and you talk about that in your paper as well. In your methods.

BSR01 31:53
Yeah, right.

Sara Mannheimer 31:54
Yeah, yeah. Okay.

BSR01 31:58
Um, yeah, I think I think I've spoken about pretty much everything about this data.

Sara Mannheimer 32:09
Okay. This is really great. Thank you so much. Okay. I really appreciate it. Well, thanks so much for your time. And good luck with this project. I can't wait to read the next steps. It seems so cool.

BSR01 34:14
Yeah, thanks. Yeah. Good luck with the rest of your interviews.

Sara Mannheimer 34:18
Thanks so much.

BSR01 34:19
Right, bye.