BSR06_transcript_deidentified

SUMMARY KEYWORDS
tweets, data, twitter, people, gathered, publish, hurricane, disaster, community, specific, paper, privacy, felt, experiencing, understanding, algorithm, fingerprint, text, events

SPEAKERS
Sara Mannheimer, BSRO6

Sara Mannheimer 00:00
Right. Um, okay, so tell me about the type of research that you do and the type of data that you produce or use.

BSRO6 00:09
Yeah. So specifically around like big social data, social media data, my work is focused on understanding how communities respond to natural disasters and climate change. And we do this by, by gathering large amounts of Twitter data specifically, it's sort of before, during, and after large scale crisis events, either doing it in real time, sort of like as an event is going on, or going through and trying to retroactively pull the sort of Twitter information in specific geographic areas or related to specific keywords. And then a lot of what I do is designing algorithms are sort of like data science methods to parse through this data, and try to correlate or in some way predict how communities are going to respond to these disasters, using the sort of discussions that we see on social media. So a lot of it is aiming to get to a more socially derived understanding of a resilient community. So it's not just like, how many power outages did you have it's, can we learn something about the dynamics of the community, based on what people are saying on social media in an effort to try to predict or to try to improve this for, for the future. We can figure what makes this community resilient, maybe I can bottle that and give it to this community to make them resilient to something else, or the same disaster, etc.

Sara Mannheimer 01:28
Hmm. Oh, that's so cool. All right. Nice. Can you please describe, so then, in my email to you originally, I'd ask you to come up with a specific project that you've done recently, that we can use as an example. Okay. We don't have to stick to the project the whole time. But it's nice to have some specifics when we're going through these issues. Yeah. So tell me about the example that you selected.

BSRO6 01:57
Cool. Yeah. So this is one specific paper that.... So again, at the idea, this, this paper was sort of like motivated by a particular like algorithm and data processing, you know, series of steps that I developed through my PhD. And it's, the idea is that a community, the aim is to try to parse out, there's a lot of, you can get a lot of information out of Twitter, there's a lot of very, like, not rich and a lot of very rich data. And so we're trying to look for just additional ways to gather information from these from just a large body of tweets. And so this was, this data processing algorithm, this umm... it's called resilience fingerprints, to like just to make this like terminology easier, is derived from a more social scientific understanding of what makes a community resilient. So sociological ideas of how people might communicate with one another or relate to one another. And essentially, we break down this concept of resilience into five sort of categories. And this is like social ties to one another, maybe more

institutional, but still social ties, things like churches, or like aid groups, things, ways that people interact in a community that are in a more structured way. There's like, of course, like institutions, police, fire hospitals, there's physical infrastructure, there's, there's roads, there's right the things that we built in our environment. And then there is health and like quality of life, so the actual healthcare outcomes that people are experiencing. And essentially, what we we do in this algorithm is, is look for, for keywords or sort of terms related to these different categories, and how they relate or they show up together. So like one example is someone tweets, 'I have run out of power, relying on neighbors for flashlights and batteries.' That's got an infrastructure component and the like, power is out. And it's got a social component, and like I'm relying on my neighbor, so we're, we're looking for those sorts of ties. And essentially, what this does is those ties we become, I'm looking to see how often does this category in this category line up with one another. And we call that a fingerprint. And so in this study, we then calculated these fingerprints for the 15 different sort of like large crisis events. These are things like hurricanes, these are earthquakes, a couple political events, like the shootings in Charlottesville a few years ago, or there was some Irish political unrest when they tried to put a constitutional ban on abortions. We looked at the sort of Twitter discourse around that, and the Twitter discourse around Brexit, and we calculated these fingerprints to try to tie these resilience components to specific attributes or specific elements of the disasters themselves. So we see things like in a hurricane, people talk a lot about infrastructure. In an earthquake, people talk a lot about infrastructure, but not quite as much. In a hurricane, people talk more about social ties than in an earthquake which is more political, and looking for these sorts of relationships to try to then understand, like what are the things I should be paying attention to, in a particular disaster. So that's sort of like... the flavor of it is trying to use the social media data in this rich, interpersonal textual communication that's happening online, to inform a better understanding of what parts of my community are being stressed or are being utilized during some sort of a crisis.

Sara Mannheimer 05:23
And you're in engineering, right? And so, in engineering, like, do you, how intensive is the social science training? Or is it more... Is it like a mixture of both social and sort of hard science training?

BSRO6 05:38
It is, there is not a lot of social science training. This is something that specifically we reached out to collaborators from a communications department, or communications background to help do a better job of understanding, to help us understand how to process information or to have a--I'm trying to remember the, the terminology, make sure I'm being correct with this-- to develop a cognitive processing model, which describes, the aim was to say someone experiences an event, then they tweet something, we wanted to have a sense of how they got from an experience to a tweet. And so we interacted with and used a lot of expertise from some people in communications to try to have a better sense of it. As an engineer, I would just, that's something that would totally get washed away. And so we really wanted to make sure that, that was grounded in a better sort of like communication or sociological theory.

Sara Mannheimer 06:35
Okay, so you use a collaboration to do that?

BSRO6 06:38
Yes, yes, exactly.

Sara Mannheimer 06:40
Perfect. Cool. Um, what was your data collection method? Did you use an, the Twitter API?

BSRO6 06:46
So we, I used hydrated... so I used a bunch of existing tweet datasets. So I was able to basically just scrape anything I could find off of the internet, someone had pre-curated a list of tweet IDs, and then did the process of turning those IDs back into the original tweet content. So all of these datasets were pre-gathered by other people. And that gave us a larger temporal horizon. So we can look at events in different parts of the world, we can look at events over a longer period of time, rather than trying to like query things in real time.

Sara Mannheimer 07:19
Okay, great. So you reuse other people's social media data?

BSRO6 07:24
Yeah, exactly. All, and I think all of it. Yes. All of it was reused.

Sara Mannheimer 07:28
How do you find that? Would you look for projects that had done similar studies? Or...?

BSRO6 07:33
Yeah, so a lot of it was from, it was easy enough to find like a specific....Hurricane Sandy was a hurricane where there was a bunch of, and I'm sure you're familiar with this, Twitter changed their policies to having like before, I think 2014, your coor- your physical GPS coordinates were attached to every tweet by default. In 2014, it became an opt in thing. So it's sort of like neutered that as like a, as a research data source. Hurricane Sandy was like the last major natural disaster to have occurred before that. So in this, like, algorithms for processing Twitter space, there's a ton of studies that do this on Hurricane Sandy, because you have GPS coordinates for every tweet location through Hurricane Sandy. So this was like a starting point. There's like one paper specifically that is cited as like the definitive Hurricane Sandy data source. And so we were able to look through like downstream citations of that to see like, who else had cited that or what did they cite? And that started to like, branch out into a couple other different repositories, but it was all looking through other papers. And then people just fortunately, publishing their data along with it or maybe sent them an email to try to get a data set. So everything that we were using was previously attached to another paper in some way.

Sara Mannheimer 08:51
And there's like some construction going on outside. Are you hearing background noise or?

BSRO6 08:56
No, not at all. Good mic.

Sara Mannheimer 08:59
Okay, good. Yeah, they're redoing the loading dock and they're like pounding concrete. Great. All right. Was this grant funded, like a situation where you had a data management plan or some other like requirements for how you handled data?

BSRO6 09:15
There were no requirements for how we handled this. This was like a sort of a pilot study that was being done to provide some preliminary results that were going to go into a grant. But all of this was, was handled just... the restrictions, the only restrictions we had on like data handling and management

were governed by Twitter's private policy and the the things you sign up for when you have a Twitter developer account or do any of those things.

Sara Mannheimer 09:40
Yeah. Okay. And then, did you publish data from your example? Or did you just point to your sources?

BSRO6 09:48
I have them all compiled and it is linked in the paper, but specifically requested that any any of, the term I think it was like any use of these other datasets should be directed at their appropriate sources, right? Like I put them all here for convenience. But if you're going to do anything with this, cite—notionally, right, if you're going to do anything with this, cite the original ones don't cite this one as the source for that. Here's the CSV file, but cite the other, the other people when you, if you want to use this down the line.

Sara Mannheimer 10:16
Nice, okay. And did you publish in a repository or as supplementary materials with your paper?

BSRO6 10:23
Yeah, it's a, it's in like, this was done at [university]. And we have like, the libraries have a data repository where you can put a DOI on it, and easily share it. So it's, it's called the [institutional data repository name]. And it was a pretty easy step to do to just hold this data sort of in addition to the paper.

Sara Mannheimer 10:41
Perfect. Okay, great. So we'll get started with these six issues. Yeah. The first one is context. And I have like a little paragraph to help describe what I mean by context. So when we collect social data from social media platforms, just like in traditional spaces, when we used to collect data context is important. However, the context of a social media post could be absent or difficult to understand. Social media posts are short pieces of text or little images or videos that are taken out of the larger context of somebody's public and personal life. And then the out of context effect can be compounded when data are collected at a large scale. So this is kind of the issue that I'm thinking about here. Can you tell me about a time if there were any during your research when you consider the issue of maintaining or understanding the Twitter, the tweets context? So like contextual information about the community where you collected the data or information about individual respondents?

BSRO6 11:51
Yeah, absolutely. And so I can think of two specific examples that I can touch on. One is a difference between like, languages, the, some of the default Twitter settings for, for just like scraping data from if you're using certain software to just scrape Twitter data, defaults to English tweets. And in somewhere like New York, that's, that leaves out a lot of people, in somewhere like Miami, that leaves out a lot of people. So this was something we were very cognizant of, was not necessarily...We ended up just doing sort of like systematic double checks to if I went back and then queried for the same period of time, how many tweets in this geographic area, how many tweets are English versus non-English, and making sure that our samples that we gathered, sort of, were not if the tweets in this area were 90%, non English, and we were using a 90% English data set? Right? There's there's a disconnect there. So one was looking at the context of the place with specifically focusing on language. And the other this is something we noticed pretty early is that there's a big difference between pulling one tweet from a person who has maybe said the key word Hurricane Sandy. And so they've ended up in a data set. What we ended up doing for this was not for this particular study that I'm talking about, this is another

one. So this is moving away from that original project. But this is a time this has happened. Is, is it turns out to work a lot better, we have way more complete, or way more representative data to find the people and then go try to retrieve their series of tweets. So I might be a person that's tweeted 100 times during Hurricane Sandy, but only once did I say the word Hurricane Sandy. And that would be the only tweet that would show up in a Hurricane Sandy database. So what we found to work a lot better was to look for the users who would show up in this database and try to pull as many of their tweets to that time period as possible. Exactly to provide the context and not just get their one out of context tweet that might show up in a data set.

Sara Mannheimer 13:56
So did you collect additional data in, in addition to what you had pulled from other...?

BSRO6 14:02
Yeah, so that was this was again, for a, for a follow up study. But yeah, we did that exactly was to take the original set of tweets, which you know, may be collected via various means, but they're all somehow related to a disaster. And if a person should, instead of looking at the data as a measure of content, if a person showed up in there, then we would go and actually pull all of their additional data so we could see those people over time.

Sara Mannheimer 14:43
Okay. So let's see. Did you consult with anybody or consider other research projects or for the literature like how did you come up with, how did you identify these issues? And then how did you come up with your strategies for alleviating the issues?

BSRO6 15:03
So I think maybe it's a little bit like a like a lucky guess the, the kind of, in this studying this idea of community resilience is incredibly place specific. And that's like if you're going to do an engineer— you know, a very technical engineering study of the resilience of a community, it is all contextualized by the place, the place specificity is incredibly important. So this was something that was like we knew was going if we're studying community resilience, because that is so largely discussed in this field of community resilience, that that was the reason why we went to this is something that needed to be addressed. Truth be told, if that was not something that was like, present in this research domain to begin with, I'm not sure how that would have, like, come up. There, there wasn't like a specific set of guidelines that we were looking at, or something they would say, you know, and you need to check the language, because that's, that's an important context. It was very much that, you know, I knew if I'm studying community resilience, I need to have the context. And so in this domain, that was important.

Sara Mannheimer 16:04
Yeah. So you were using, like, knowledge that you had gathered through other studies? And from your just knowledge of the domain?

BSRO6 16:10
Yeah, yeah, sort of existing work in the literature, which says, look, the way that this community responds will be different than this, you have to consider that community level that place specificity that that context is important.

Sara Mannheimer 16:20
And did you write about this in your paper, too? Did you add, like, when you, did you talk through these ideas of doing double checks for language being spoken? And looking at individual users?

BSRO6 16:35
Yeah, and maybe this, I don't, I don't know, this can remain on a recording, but maybe like, there's a paper under review right now that I'm just trying to be sensitive to like, discussion points, can...

Sara Mannheimer 16:49
I'll make a note.

BSRO6 16:53
But to to sort of like back this up, we actually ended up doing a complete study based just on, we took hurricane, there's multiple different ways you can gather tweets, and we just took a few different ways, so I can get tweets that are based on keywords, they say Hurricane Harvey, or maybe they were in the geographical region of Hurricane Harvey, and then just did a bunch of analyses that would be sort of typical of the social media analysis. Let's see how many people retweet and how many people post images and gifs and how many, like individuals overlap in these datasets. And they're entirely different. I mean, there's, they're, in almost every analysis we did, the way we gathered, the tweets was incredibly important. There're like 1% of users showed up in both. There's no overlap. I mean, every analysis we did was like, fundamentally different between the different ways that we gathered the tweets. So this was like, yeah. This was the sort of like, follow on that, you know, we realized in the course of this paper that like, yeah, this is this is a huge component of this, but probably deserves a little bit more attention than like, a discussion point. So let's try to like actually codified this in, like a more analytical way.

Sara Mannheimer 18:26
That's great. Cool. All right. Um, let's move to data quality. Can you tell me a time during this research, when you considered the issue of data quality? So like, missing data? Or if there were busts in the data set, or the quality of other people's Twitter data collection, stuff like that?

BSRO6 18:47
Yeah, yeah. I mean, this is like, this is like, I would describe this as the fundamental issue with like gathering Twitter data, at least from the perspective of like a data science from a very data science perspective, it's, it's very messy compared to other things. I'm just like a normal like good old fashioned data scientist. This is like as messy and as unclean and as unreliable as data sources come. So it definitely was a concern through all of it, specifically, so things that were like known issues that we tried to deal with were that when you're rehydrating these tweets over time, you have, there's a, there's a loss, people either delete tweets, or they make their accounts private. And so there's this, there's a pretty like regular, just like decay of this data over time. And there is a particular, there's a paper that tested this specifically. And they, they basically just wait, they gathered a bunch of tweets, both like related to specific events, and then sort of just a broad sample of Twitter. And then five years later, they tried to go based on the idea, they tried to re, re-access the same data and they found I think it was about 75% of the tweets were still there. So they in five years, they lost 25% of their data. But by any means that they could find the remaining data was still a representative sub sample. So, so when you're looking at users, when you're looking at geographic distribution, anything that they were interested in testing, there was, it was a pretty nice sub sample. So it seemed as if that decay or if that that loss was pretty evenly distributed. And so that was pretty reassuring. We had losses, I think, in about the like, at somewhere between 85% on average of the tweets that I went to go access ended up being accessed. So that was one was just the some of it's just straight up missing. When you go to re-access it further on down the road. We did test for bots and remove bots using I think Indiana University has like a, an API for check you, you run a user through this, and it checks to see how many bots are in

there. We, we attempted to remove them, and ended up not removing them in all circumstances. Because there was a lot of situational awareness type information that was present. Local weather stations will tweet the status of hurricanes or local news organizations will retweet a news headline, and that tends to be flagged as a bot. It's an automated Twitter account, but it is still providing situational awareness type information. So we did test for bots and reported the percentage of bots in them but didn't remove those. Because in certain circumstances that, our judgment call in this case was that that information is more useful and does provide context for what's going on, versus removing them where we sort of have this this gap that we have found.

Sara Mannheimer 21:59
Again, did you like, what were your sources for understanding quality and figuring out the strategies for dealing with it? And then also can you... And yeah, I'll stop there.

BSRO6 22:11
Okay. So the primary sources for [strategies for addressing quality] were existing papers was was looking to the literature specifically on gathering social media data around natural disasters and natural hazards. There's a, there's a reasonable enough amount of publications in that space that these issues, we were able to find these issues through literature reviews, and know that these were going to be coming. That was the primary source of trying to figure those out.

Sara Mannheimer 22:41
Okay, and then you did you document this in your paper as well? Yeah.

BSRO6 22:47
I think we have an entire appendix that's just like, here's what happened in the data collection process of things. Because there's just so many little steps like that. That are, our goal was to make it as repeatable as possible. I mean, all the code is published. But so you would understand if you would see if you read through the code, like, oh, you're getting rid of all of these things, or you're getting rid of all of these things. But our aim was to provide a more detailed justification of that, should someone be interested in replicating.

Sara Mannheimer 23:16
Yeah. And so ultimately, you were like, you felt that these things you did to help to support data quality were enough? Like, yeah,

BSRO6 23:26
Yeah. But very much a judgment call, like the very much a judgment call internally as to whether or not that was enough. Our.... Yeah, that was our eventual like result was to essentially look at by compute this, like analytical fingerprint for a given disaster, we we're interested in the sort of similarity of these fingerprints over time, or among different events. So is the fingerprint before different than the fingerprint after a hurricane? Or is this is a hurricane fingerprint, similar to an earthquake fingerprint? And so along the lines, we were sort of like checking to see the impact that it might have on the similarity of these fingerprints. And so if we removed the bots, does the fingerprint similarity remain the same? If we add them, so then if it doesn't, if removing them doesn't fundamentally change this, we're going to include them because then we think it does provide richer information. So a lot of it was tested sort of empirically against what we were aiming to see. Okay. So similar with some of the like, text removing, or text cleaning is another like obviously a huge part of this, do you leave emojis in, do I get rid of the word amp or like RT because this though the presence of those is, you know, indicative, it's comes in the text, but it may not necessarily be indicative of providing additional information other than

that, it's a retweet. Do you filter out swear words? So a lot of this was was very much done to see okay, if I get rid of emojis do I have a less noise and more signals sort of in my result, and we could test things that way which was, which was good but the the when it was good enough was was totally a judgment call.

Sara Mannheimer 25:06
And did you discuss that with your research team?

BSRO6 25:10
Yeah, this was all all of us would sit there and look at these plots and say, I think it actually makes sense to do this one.

Sara Mannheimer 25:15
Okay. Okay. All right, let's go to data comparability. So during this research, you did combine multiple datasets, right, that you had found elsewhere? How, like what issues arose with combining of data?

BSRO6 25:38
That's a really good question. It's, one was the sort of scale or the scope. So it's unclear even now, where these differences in scope came, there are some natural differences in this scope. If you are one of the events we were looking at was an earthquake in Nepal. Twitter adoption is just, you know, orders of magnitude lower in in Nepal than it is in the United States, or in even in the United States, something happening. There's a river in Oklahoma, that floods, the Red River in Oklahoma floods pretty regularly. The Twitter adoption there in rural Oklahoma is a lot less than New York. And so Hurricane Sandy has this like, it's easy enough to hypothesize that the average person impacted by Hurricane Sandy is more likely to be on Twitter, than the average person in Nepal who's experiencing an earthquake. So one was this this concept of the place bed, the the context or the place specificity of the disasters we were actually studying. What we were finding is that, by and large people who were tweeting about an earthquake in Nepal, or people who were tweeting about the Ebola crisis going on in Africa, were not actively experiencing the Ebola crisis in Africa, they were commenting on it, a higher proportion of people who were tweeting about Hurricane Sandy, were actively experiencing Hurricane Sandy. So that was a very difficult attribute to deal with in this data processing. Because we're essentially we have datasets that, by their numbers, by their like structure are identical. That's, it's just tweets that have been gathered through a given period of time, but in their meaning, or in in what they sort of are representing the types of discussion or online discourse they're capturing are fundamentally different from one another. So we ended up getting rid of a lot of those that that were not as comparable so that the comparability was a lot about what we thought was being captured by the tweets as opposed to the sort of technical interoperability of the data sets.

Sara Mannheimer 27:42
Yeah. How did you identify which, which tweets were commenting on the disaster rather than experiencing it?

BSRO6 27:53
A lot of it was just looking at like Twitter's annual reports, they have a pretty good description of their monthly active users by country by time period. That's, that's pretty easy to find information. So we would go through and just look and say, how many monthly active what's the percentage of the population of people in Nepal in 2009, that have Twitter and it's, so let's, that's probably not people experiencing a disaster, we'll strike that. So we used like, a proxy was, was how much Twitter was used in a given area in a given country at roughly the period of time when the disaster was happening.

Sara Mannheimer 28:34
Cool. Okay. One thing I meant to ask you before was did with quality, I guess, in comparability too, but like, did you encounter when you were trying to reuse these lists of tweet IDs that you had found? Did you encounter any, like problems with the way that other people had documented their data or missing information about the data that made it more difficult to use?

BSRO6 28:56
Yeah. There, this was something that is, it's it's not easy. But it is possible to do after the fact to go through this list of tweets. And I, it's not hard to see that everyone contains a keyword. And so this was probably gathered because they just look for everything matching a key word, or some of these are just total non sequiturs that aren't related to this at all. But if you go to the profile of the person, you can see that they are located in Houston during Hurricane Harvey. So we were able to sort of intuit that over time that they may be gathered in different ways. But yes, there were no clear reporting guidelines, there's by no means a standard as to it. It took us having to go through and sort of manually double check all of these to make sure that we could find a reasonable have a reasonable guess as to how this data was gathered or what the sort of the sampling rates are, how much data they got, we had to go through and figure that out after the fact there was no documentation for it and that was a real sore spot, so to speak.

Sara Mannheimer 30:01
Yeah, were any of those, that data you were using, did you find it in like specific data repositories like....

BSRO6 30:09
Most of them were, like, attached to githubs that were like, attached to papers. And I think in three or four cases, we were able, there's a program called Social Feed Manager, which is like a software for scraping tweets, that does a pretty good job. And that has what I would, I don't know if archival quality is like the right term, I'm not sure if that term has specific meaning, but it has very high level very detailed reports as to what filters were being used all of the query attributes that were being applied to it to pull tweets out. So those were the like, the best of the best, because I know exactly what was being used to do that.

Sara Mannheimer 30:50
That was developed by librarians.

BSRO6 30:54
Yeah, it definitely. It was it was excellent for like our purposes to be like, 'Okay, I know exactly what's going on here.' Where especially older tweet data sets, it was someone very, a lot of them felt like it was someone who had written their own custom code to like, just go, just go try to get as many of these tweets as we possibly can. Yeah. So yeah, that was really specifically those were like the best examples, but by no means was that a universal experience.

Sara Mannheimer 31:21
Okay, great. Very interesting. All right, let's move to informed consent. So can you tell me about a time during your research when you considered informed consent?

BSRO6 31:32
Yeah, I mean, it's like a little creepy. Like, when you first like open one of these, you take the tweet IDs, you hydrate them, and you're like, 'Oh, this is just a person and you live here, and you just tweeted

about your house.' And that's just on the internet now. Ultimately, the, like any concerns we had does, from like, a moral sort of invasive privacy standpoint, it feels a little icky. But in terms of like, you know, what actual regulations are there, we were again sort of leaning on the Twitter private policies, the policies, the terms of services that are Twitter's Terms of Service and how they govern the use of these developer accounts that you have to have to access this data. That was what we kept going back to to say, 'Okay. According to these rules, it is okay for me to publish this data.' In no cases did we use like actual tweets, content, just sort of, like, out of respectfulness didn't feel like the right thing to do to just publish the content of a tweet, despite it being public and being aboveboard to do. But that was it was the terms of service that sort of like governed that.

Sara Mannheimer 32:40
Tell me more about that, like feeling that it wasn't right to publish. Where did that come from? And then how did you decide among yourself that that was going to be like a part of an internal policy that you made?

BSRO6 32:53
That's a good question. I don't know like where specifically it came from, other than to say, like, you just doing some sanity checks, just sort of like, 'Alright, I, I, you know, little grad student. And I've, I've run my code overnight, and I wake up the next morning, and I go into the office, and I like, Oh, look at all these tweets that I have. And I can just kind of scroll through them. I'm like, Look, let's see what I just pulled off the internet.' It is very invasive. I mean, people would be talking, especially in the data sets that were gathered in a more sort of dragnet fashion where it was maybe any tweet in a geographic area, or any tweet in a given time period that was in a geographic area versus tweets that were containing a keyword for a disaster. Those tended to be pretty intuitive, and pretty standard, you can you can, those were a lot more predictable. But if it was just everyone in this region during a disaster, you got a lot of things that weren't disaster related, despite them being public. And so with the feeling arose from seeing a lot of the things that from just experiencing, or seeing a lot of the things that people would post, and it is not relevant to what we're doing, per se, this is not like necessarily informing our analysis, but it is data that shows up as a part of it. But no, nothing that we, everything is essentially de, de-identified or anonymized in our algorithm. So we pull text, but it's no, it's never attached to a specific user once the processing is done. So when we're actually processing or we're actually computing things, all of that is done on, on a large corpus of text, not on a specific user's individual tweet. So when we were publishing it, there is no requirement in the slightest that we have identifiable information that we didn't need to publish specific tweets, none of that was required to improve our analysis. So it's, if we do not have to this is absolutely not something we should do.

Sara Mannheimer 34:49
Okay, and so you didn't publish any, like, sample quotes or anything from...

BSRO6 34:54
Yeah, we had, you know, one which sort of described we used a couple that were essentially made up, we would just I would tweet something. And then we would run it through to show like how this the algorithm pulls out this word to mean this and this word to mean this, any such circumstances of those, I would just tweet and just do it myself. So we had one sample, but it was, you know, a purely orchestrated made up example, that was just a tweet that had been put out on the internet.

Sara Mannheimer 35:25
Okay. And was there anybody that you looked to for guidance, to come to this strategy?

BSRO6 35:34
Not particularly. I This was I think there was a team of four of us that were working on this. And so this was just an internal discussion as to sort of like what felt like the right thing to do. Our, our group and the people who were doing this had a bit of background on working with like, power company data. So again, it's like resilience of communities to disasters, we had previously together done a study, looking at power outages in response to like Hurricane Katrina. And this is like really sensitive, fully identifiable data, you have a name, you have an address you have when their power went out when the power came back on. And in circumstances like that, like the sort of standard protocols were to de-identify everything immediately and only work with anonymized or de-identified data. And so that was a bit of experience that we had had that, you know, if possible, if there's, if there is something that could in any way be considered sensitive, it makes sense to de, de-identify and go on and use that. So that I think also contributed is some sort of historical work, having done work with PII led to this idea that maybe this isn't like PII by the letter of the law, but it is PII sensitive adjacent. Yeah. And so it felt like a right thing to do not necessarily if it was governed by something.

Sara Mannheimer 36:56
Yeah. Okay. Sensitive adjacent, I like that. Okay, and the next issue is privacy and confidentiality. So this is closely related. But was there a time during your research when you considered the issues of privacy? How and then how did you deal with those? How did you conduct the de-identification? And how did you use the aggregation in order to support your conclusions, but while supporting privacy?

BSRO6 37:29
Yeah. So the analysis that all of the algorithms that we developed are just based on text. So our sort of privacy steps were to take this series of tweet IDs, rehydrate them. So we've got all of the information we need, just extract the, the raw text, and then that would be the only thing that would leave, we had a particular secure, we call it the secure server, it was the one that was just a workstation that was doing all of the lifting. And then anything that we would take off of that was only text. So our de, and that was our de-identification processes, then for a given disaster, I have a series of tweet texts, but it's not linked at all to an account, it's not linked at all to a person, it doesn't even have a tweet ID attached to it anymore. So that was what we, again, you know, we could have just used from, from our understanding of the terms of service, we could have used all of it, and it would have been okay. But that felt like the sort of, if we didn't need it, it wasn't important to, to bring that information along. Let's try to use as little as we can. And there was one specific instance. This was a disaster that I think we ended up not including in the final paper. But there were situations in which there would only be...It was a very small earthquake in California in like 2009, it's really in the fledgling, Twitter was super young at that point. And the data that we had was really small compared to some of the others, right, Hurricane Sandy has like 13 million tweets, I think totally it's 250 million tweets, and this had like 25,000 tweets. And so when we're looking to try to match, like maybe this tweet matches institutions and political, there were a couple situations in which there was like one tweet that matched those. And we didn't, we didn't end up including this. This was not published. We didn't we didn't write this analysis. But there was a consideration like if there is there is one tweet that matches these things. I guess in theory, you could back out which one it was. And it still felt a little identifiable in the way that like, you know, the census doesn't publish data, if it shows that there's a there's a lower limit on income. So they like if there's only one house making a given income in a given neighborhood, they won't report that so you can't go 'Oh, it's these people that make this income.' So you know, sort of in a similar vein, if it were getting to the point where there was only a few tweets that were like ending up motivating our analysis, we got rid of it. But that was another time where that was sort of like, potential questionable privacy issue.

Sara Mannheimer 40:17

And did you... Okay. And then with your data publication, you didn't, did you publish the text of the tweets, or you also just published tweet IDs?

BSRO6 40:27
We just attach tweet IDs in our sort of repository. I think that's I think that's a Twitter Terms of Service thing is I'm only allowed to share tweet IDs. And then in our paper itself, there was no text. So it's just....

Sara Mannheimer 40:41
The analysis is with the text, but then you publish the tweet IDs, and then you have your code that shows how you went through it to pull it.

BSRO6 40:48
Exactly. I'm not actually I'm not sure if we published the hy-, like the code that takes the tweet IDs and hydrates the tweet IDs. But we still think it starts with let's assume you have a large bit of text, and then like our code that we've shared starts there.

Sara Mannheimer 41:01
Nice. Okay. Did you...So what did you think about what the Twitter users' expectations of privacy were, you know, thinking about the terms of service and like how the users understand those?

BSRO6 41:20
No, truth be told, and probably should have, I think this is something that I like to think I'd be more like aware of, or mindful of now. I just think that we did this work in 2017, into 2018. And so it is, like, in my personal life, privacy has become like a larger concern over time. So my hope would be now that that would be the case. But I think at the time, that was not the expectation of what a user was hoping their data would be used for, really didn't didn't play into it.

Sara Mannheimer 41:53
All right. I know, it's funny how I feel like our perceptions about online presence and social media at change, like every three months. There's so much... Yeah.

BSRO6 42:06
I remember being like a teenager, and like sneaking out of the house and going to a party. And just like the next day, like don't post that on Facebook, right? Like, you know, 'Don't tag me in that picture on Facebook,' and like, and thinking about then versus now, where if someone did that, that would just be that would be socially like a really bad, you would never do that you would never do that. Because they would have social repercussions. And as a society, we've learned like, now you just don't tag embarrassing pictures of your friends when that we had to learn that at a certain point. That was like not something that we had figured out. So yeah, I think I hope that my like user understanding of privacy has grown. But hopefully everybody's has.

Sara Mannheimer 42:45
Well. And my, I hope that eventually we have like policies that can govern this stuff, because it's a little ridiculous that you were relying on individual researchers like you to be like, you know, it didn't feel right to do that. So, like the IRB should govern that or that, you know, so. I'm letting my own thoughts come through. All right. Our last question is about intellectual...Our last official question is about intellectual property. Was there a time during the process of this research when you thought about intellectual property concerns? Yeah.

BSRO6 43:26
I don't think so.

Sara Mannheimer 43:31
For example, when you were reading the terms of service, Twitter's Terms of Service, like, it sounds like you just really went through those and followed those to the letter of their law.

BSRO6 43:42
Yeah, yeah. What, what I know, is a concern, or again, this is sort of like the, there are people so this was what would have been what would have been a concern with regards to intellectual property is like the, the nuances or the specifics of the algorithms that we developed. So again, for example, this group had previously done work trying to predict the number of power outages in a given place. There's now a lot of companies that sell that as that's their, the algorithms, they will predict power outages, and they will sell that information to cities, and the, the underlying tools they use to predict that are their IP. And so our concern, I think, was like if this, let's say I could make a really great prediction with social media data about how the community would respond to disaster XYZ, that the process of making that prediction was our concern for IP. And so you know, maybe someone was going to steal our algorithm and do something with it or use it to they could use that as the basis of a company that they could try to sell something. But that I think is mostly covered by University IP policies. If you, I'm a student at this university working with a professor at this university, I think the university, my understanding was that my university just owned anything that I created while I was working on there. And it would be the university who would work on those sorts of like, IP disputes on my behalf. With regards to like the actual content from Twitter, now, there was nothing specific that was like, privacy was a much larger concern than intellectual property.

Sara Mannheimer 45:34
Yeah. Okay. Good. Great. So the last question is, is there, were there any other issues or challenges that came up during the research that I haven't asked you about?

BSRO6 45:48
That's a really good question. I think the largest challenge is, is more of like a, like an academic and sort of process one, which we've already touched on a little bit, but it is like the, the interface between having the sort of technical and like computational know how to access all of this data, process all of this data in a way, which has some level of integrity, then ultimately, you're, you're ending up with a data set that is very much the result of human processes. And so it was like, I had the expertise to go on the internet and pull all of this data up. But I have no expertise on how on theoretical or sort of conceptual understandings of how people process information. And it was tough to find people who were to, it was a tough collaborative effort to try to find people who could be at this intersection in to be programmatic enough to pull 150 million tweets from Twitter, the Venn diagram of the people who can do that, and the people who have like, firm social scientist training and understand what this data means is vanishingly small. And so it was a lot of collaboration and a lot of discussion to try to create a team that could sort of balance both of those, I think that was a very big step, because it's very easy to say, 'Well, I pulled all this data out, I'm sure I can find something to do with it.' But I have no idea if that represents what any sort of way that a human might actually think or act. And so I need to involve that expertise as well. So that the technical hurdle was limiting. So you had to have the technical skills. But that did not necessarily mean you had the formal training in a more social scientific side, to understand what that data meant? That was a huge challenge.

Sara Mannheimer 47:47

How did you find the communication faculty who you reached out to?

BSRO6 47:52
Oh, pure luck. I had a, I had a committee member who was like 'Oh, you should talk to this person.' And we started speaking like, wow, this is like really great that we ended up here. It was an issue that we knew we were going to come across. And so eventually, we're going to start just soliciting help. But that was yeah, it was luck that we got connected with this person in particular, and certainly very fortuitous that that happened.

Sara Mannheimer
And did they end up being a co-author or an advisor?

BSRO6
They were yeah, like sort of more of an informal mentor. I think they're mentioned as an acknowledgement. But they were because the nature of the relationship was very much 'Hey, I have these things. Does this make sense? Can you like, you know, sanity checks some of these?' Is it okay, well, here's maybe some papers you should read. Here's maybe some like, here's a textbook that's going to help you like catch up to speed on this. More so than someone being like directly involved in sort of contributing in an authorship way.

Sara Mannheimer 48:53
Awesome. Cool. Well, thank you so much. This was so fascinating. It's gonna be a great addition to my, my research, so I really appreciate it.

BSRO6 49:48
Yeah, absolutely. So it was great to meet you.

Sara Mannheimer 49:50
You too. Have a good day.

BSRO6 49:52
Yeah, you do the same.