GESIS Guides to Digital Behavioral Data #19

Expert Insights into Digital Behavioral Data for Health Analytics

An Interview with Sanja Šćepanović

Publication date: April 24, 2024

Health is both a major personal and a societal concern. At the same time, it is a sensitive issue, and health data come with a lot of challenges on all levels: be it data privacy and research ethics or the problems with getting reliable information on people's health in the first place. Digital behavioral data can offer new insights for public and individual health, especially when it comes to studying mental health conditions, digital epidemiology, using new types of sensor data, or connecting multiple data sources.

Sanja Šćepanović is a senior research scientist at Bell Labs, Cambridge in the Social Dynamics group led by Daniele Quercia. She completed her PhD at Aalto University in the Data Mining group with Prof. Aristides Gionis. She researches health through social networks and mobile sensing, and urban outcomes using remote sensing.

The interview was conducted by Indira Sen and Leon Fröhling on July 18, 2023. We met Sanja during the 9th International Conference on Computational Social Science (IC2S2-23) in Copenhagen. The interview has been edited for clarity and length.

Keywords: health data, public health, social media data, digital trace data, digital epidemiology, mental health, individual level data, wearables, sensor data, green areas

GESIS: Hello Sanja, thank you for giving this interview and providing insights into your area of work. First, a very general question: What are you working on at the moment? Especially in the context of analyzing individual and population level physical and mental health with digital trace data?

Sanja Šćepanović: I work a lot on different health topics and social media health discussions, some of which I will present during the IC2S2. We developed a method where you can extract any symptom mention from text, for which we have several applications. One was to study stress across companies and another one was to study public health across the United States. I am currently expanding my work on public health applications using various public health data sources, for example, prescriptions in the UK in relation to satellite environmental indices [1].

> *The problem is not the data anymore, it is how we make sense of it.*

GESIS: How did you first enter this area?

Sanja Šćepanović: My master's was in a very different area, in cybersecurity. But then I found myself in California, attending a lecture byJure Leskovec. He was presenting on computational social research, calling it social network analysis at that time. The main story that Jure was telling was that for the first time in history, we are collecting so much data. Before, social scientists, psychologists, they all had to spend so much time to collect data, mostly by interviewing people. Now all this data is at the tip of our fingers. The problem is not the data anymore, it is how we make sense of it. That for me was convincing, that this would be one great way to do science nowadays, to have the power to make sense out of this new data, big or small.

> *[Mental health] is one of the main issues nowadays. Social media is a good reflector in some way.*

GESIS: And then, specifically on digital health?

Sanja Šćepanović: Then I worked on various topics which were not particularly health related until I joined Daniele Quercia's team. When I joined, there was the question on what we should start working on. There was an open project because a person left. This project was a little bit about what I mentioned earlier, about how we extract and parse health discussions. It was in a very early stage, and that was interesting to me. But at the time I did not imagine that it would draw me in so much, I thought it is just one of the projects as I had done them before. In the process of working with the project and the data, I realized that it is a very interesting and useful application. Before starting the project, I did have interest in mental health in particular, which I think is one of the main issues nowadays [2, 3, 4]. Social media is a good reflector in some way. This one started spontaneously; it was this big project, and then, once I executed it well [5], it led to many applications. It started shaping as a research topic of mine.

GESIS: Were there any other major inflection points? Or points where a new method or a new understanding for you popped up that informed your path?

Sanja Šćepanović: Yes. And I think not just for me, but for all of us. The development of AI is so significant! When I was working on my PhD, for example, I was trying to analyze the homophily and semantic relatedness of discussions on Twitter. To be able to do this, I had to download the entirety of Wikipedia and create an embedding space, because there was only an old method before all this embedding deep-learning came. There is an old method by a researcher at Google, Evgeniy Gabrilovich which was to download Wikipedia, then do tf–idf analysis on the articles to find how related some words are [6] then build your own embeddings, and finally apply it on your own data, for me Twitter. I did this, which was an enormous effort. Fast forward three years and it is two lines of code with SBERT now.

> *There are inflection points where you realize that the developments of technology are so tremendous that it is important to have good questions.*

There are inflection points where you realize that the developments of technology are so tremendous that it is important to have good questions. If you are into applications, having good questions might even be more important than honing your technical skills, because that is getting developed on the side, if technical and method development is not your main area.

GESIS: That is very good advice. Going a little deeper into the topic, you already touched on the presence of new forms of data, and I want to go back to that. You have analyzed a lot of health-related conditions and topics like stress or dreams, could you tell us a bit more in detail why digital trace data is so interesting and valuable for this type of measurement?

Sanja Šćepanović: I am not a hundred per cent convinced the we found the best usage of digital data for health yet. If we could work with prescription data on an individual level, if we could interview all people, or be using doctors’ records, all of that would be better. What we do with digital traces is a proxy. But I do think, especially for mental health, that social media can give us even more, because people are not always truthful when they talk to their doctors. Right now, for example, I am working with prescription data. You do have biases, where maybe some population is more aware that they can use mental health medications; if they assume they are good, they will use them more. This is not because they have more mental problems, but just because they are more aware of the medical options. So even those data come with biases, and I think social media there can reveal things that even doctors do not understand.

The other thing which is useful, but I still do not know how we can do it, is to see health from people’s perspective, not just from the medical experts’ perspective. Our work on symptoms revealed that medical taxonomies define conditions with some common signs that doctors think define a condition, but people actually talk about other symptoms, so doctors and patients do not express it in the same way. This might be useful in some ways, but I do not have a concrete idea on how yet.

GESIS: You already touched upon some of these proxies that you use to approximate the real thing that you are interested in measuring. Maybe you could expand on the most frequently used types of digital behaviorial data for studying health related topics. And also on the methods you use most frequently or that you are most excited about.

Sanja Šćepanović: I should mention data from wearables, even though that is not directly my main area. I do have two pieces on wearables [7, 8].Because this data is usually individual level data, it is going to give you more information than social media data alone can. That is my perspective now. For types of wearables I think about smartwatches as well as more unusual types of sensors, like an abdominal body sensor. You can sense sounds, for example, coughs for COVID, audio data, or PPG signals and so on.

> *Another important thing is how we can have these methods trustworthy and trusted by doctors.*

The other part is about methods. Now we have conferences devoted to machine learning for health, where methods are developed particularly for health data (e.g., https://chilconference.org/) Another huge part is electronic health record data from collaborations with hospitals; these notes, written by doctors, are just like social media data, free-form text. They are just also natural language processing tasks, and a big part of the challenge in that community is to transcribe these records in a unified way so that we can analyze them well. There is a lot of work there, it is not my work, and I think now this is in some way solved. Then we have radiology scans and image data. So this conference is devoted to methods and finding ways of analyzing these texts, or even radiology scans and image date, while being careful about biases, which is super-important.

Another important thing is how we can make these methods trustworthy and trusted by doctors. The definition is not clear yet, because this is subjective and personal. The research there is not going into technical definitions, but into how we can put this in practice with doctors. We had a lot of examples for epidemiology and for COVID, where social media was tracked and it was shown that you can use it for health analysis. I personally have seen very few examples of this approach still being used by public agencies. Part of the challenge is explainability. What are you really analyzing? Can you prove to me that what we are analyzing is truly what we want to track? Because if I am a public health agency and I survey people, I know what I collected, even if people did not fully and correctly answer. Those are some challenges in terms of methods for digital traces.

There are two professors at Duke University (Cynthia Rudin) and at UBC (Margo Seltzer) that are building interpretable machine learning models for tabular data [9, 10] Tabular data is actually a big part of health data. Instead of giving the doctor a model without further information, they are finding a set of all equally performing models, or models performing within a reasonable boundary, and show the doctor all these alternative models.

The doctor can then see which variables which model is using, and may choose the model accordingly. This seems to be one way of breaking this trustworthiness barrier.

GESIS: Trustworthiness is an important topic to touch on and it shapes how not just researchers, but practitioners will be using the insights of our research.

Sanja Šćepanović: One should think about this, even if you do pure research. Your goal is in the end for it to be used.

In general, such studies are of smaller scale now and cover a shorter time span; as a consequence, the validity is always a limitation and we do not know how well our findings generalize.

GESIS: Continuing with what you just mentioned, could you tell us about your experience with sensor data specifically?

Sanja Šćepanović: There are different challenges. With sensor data it is harder to collect enough data because you often need an ethical approval for your study, and then you need participants willing to participate and to consent. You can then collect data for a week or ten days maybe, but usually it is hard to collect over a longer period. My colleague Marios Constantinines is working a lot on this, and the key word they are using all the time is 'in the wild', which means: how can we make this go beyond collecting a week's time and ten participants? Can we have devices which we can give to people to wear for a year or half a year and collect? This is not solved fully yet.

There are some studies which collected a lot, like the Cambridge University study with coughing [11], because it was during the COVID Pandemic and it was easy to get participants.

In general, such studies are of smaller scale now and cover a shorter time span; as a consequence, the validity is always a limitation and we do not know how well our findings generalize. I can talk concretely about our work with the abdominal body sensor [8], where you need to put a sensor on people's stomachs. This was again not easy, you really need to find students or colleagues who want to help. Even then when it comes to analyzing, you throw out a big part of this data, because some people did not wear the sensor properly, or maybe they were not supposed to eat but they ate in the morning, or they just forgot it this day. There is a lot of data cleaning needed, and a lot of missing data to account for..

GESIS: That already points to answering the next question, which is about the current challenges and developments in studying digital health. Maybe this is related to this distorted or missing data you just mentioned?

Sanja Šćepanović: Definitely, and that is what you find in the key conferences; Machine Learning for Health (ML4H) and the Conference on Health, Inference, and Learning (CHIL). You will find in those conferences a lot of development of models for missing data or for uncertainty, and on how to deal with these uncertain results. The other set of challenges is the explainability we talked about. There is also a lot of work on trying to make these models more easily usable.

GESIS: Connecting this theme of challenges and potential opportunities back to this other very big recent development, which is the large language models that are used for pretty much every domain there is. Do you have an opinion on their role in health related topics, what sort of applications do you see, or do you think their use is rather complicating the situation?

Sanja Šćepanović: The symptom extraction work that I have spent one and a half year to develop, we could now just do it with ChatGPT. It would be pricey, but that is the only problem. In the past I had to work really hard to label my data with Mechanical Turk or other crowdsourcing platforms, to train different models and to find the best one – I would then eventually have a model that extracts the symptoms pretty well, but not perfect. I believe now we could show the same text to ChatGPT and ask it to find all the symptoms. We could even ask it to do something which we could not achieve before, like dividing it into different symptoms that map to the different conditions.

This is called concept normalization. For example, alopecia means losing hair. So if someone on Reddit says "My hair is thinning", you would want to have those two linked together. ChatGPT could actually do it, but I have not yet seen this work, and it would be quite important to evaluate this process in order to know that you can trust it. This could be applicable, but it is still expensive on a large scale, it might cost some money to process billions of Reddit posts, as we did in one project. The other set of applications that starts to appear and is being discussed is patients that used to go to Google to search are now going to ChatGPT to search. Some early works show that this apparently works well. It again brings the question of how much trust people can have in this.

GESIS: Many potentials, but many disadvantages as well. On the negative side of things, are there any kind of misconceptions or misunderstandings that people might have when studying digital health? Are there certain things that people assume about the area that might not be true, certain things that they constantly get wrong.

Sanja Šćepanović: For some time in the past, there was this idea of using people and their web searches to study population health, which was adopted even by Google. They published a paper on it and even developed a platform, but then they had to shut it down again after a very short time (Google Flue Trends [12]). The reason was that they had fine-tuned their model to work on all the searches, thinking it is picking up only flu. But – and this goes back to explainability and interpretability of the models – the machine learning model also picked up on many other factors, e.g., since flu cases mostly happen in the season of winter, the model would connect other things that happen in winter to flu, and not just the symptoms of the illness. It was a failure. It initially worked well, and then really disagreed with the official statistics.

So we should be careful with what we are capturing. What is my advice there? I do not know, it is very tricky. Maybe my advice is to really try and think if you can explain or interpret your results, and try to double- or triple-check them. Be very honest about the limitations.

Also, a lot of the work we do is on population level, ecological studies. It might then be good to always discuss the type of study, and that for any true policy implications, this needs to be picked up by individual type of studies. We might propose some hypothesis, that there is some effect on health, but this would need to be validated by individual level studies. Our studies are a first step, but they are not a final step at the moment.

> *I feel anything to deal with mental health is a bit more suited because of people being less stigmatized or maybe feeling more free to express themselves when they can do it anonymously.*

GESIS: Do you think that there are certain scenarios that are very well suited for this type of digital health analytics, versus other parts where you would say it is better to rely on the more traditional, more established ways of studying health?

Sanja Šćepanović: In general, as I said in the beginning, I feel anything that deals with mental health is a bit more suited for this digital analysis, because of people being less stigmatized or maybe feeling more free to express themselves when they can do it anonymously. This gives more space to understand people's well-being and health. But if you want to arrive at a precise understanding of some conditions, we really need additional data sources, because people may not talk about all those conditions voluntarily and unprompted.

For wearable data, I think well-being and stress is something that we should track, because these are less clearly defined by doctors. Proper diagnostics really require close collaboration with medical experts. One very important thing about these population health or across-the-country studies is that we always need to acknowledge whether or not our studies are – and I think 99 percent are – association type of studies, which means we do not find causal links. We simply cannot know if it was the environment that caused the increase in asthma percentage we found, or if it was something else. Causal methods, once they are suitable for our data, will be important.

GESIS: You already mentioned ethics approval, but given that this is quite a sensitive topic, do you have any suggestions or advice for people who are just getting started with this type of work? What type of constraints are there or what type of things should they do to remain ethical and take good care of the research subjects and their data?

Sanja Šćepanović: I think ethical approval, for the studies where it is needed, which is many of those using wearable, should really be thought of from the beginning, already when designing the study. If you are not experienced, it might be good to write an early email or have some conversation with your university's ethical board, just so that they are able to steer you early on, showing you what is completely impossible.

The other thing is to really take care of not hurting or harming your participants, which can happen in those studies. This includes being really, really careful about the protection of the data, for instance not sharing them unless it is strictly necessary. And also being transparent with the people about this.

There was a recent study at CHI [13], a conference in the field, and I was expecting that this type of conference would be adhering to these standards to the most, but a big percentage of papers did not have any consent forms. And a big percentage of the papers did not adhere to the principles of openness and transparency fully. This second part, which is somehow related, is sharing your code. You often cannot share data, but what you can and should share is your code.

> *For the data there are frameworks which are called Datasheets for Datasets.*

GESIS: I was about to follow up on the topic of data sharing; you already said that this is a very difficult issue, and you probably do not share datasets for all of the things that you do. But if you do, are there best practices or procedures people should follow if they work in that area and with that type of data?

Sanja Šćepanović: I think this is a very good question. Obviously, do not share data exposing individuals. Even when you share Twitter or Reddit data, it would be good to be careful to minimize the possibility of linking. Because this is already all public data, you could share the data, especially in some post-processed forms. You might still want to check with your supervisors or board. For the wearable data, there are already some datasets which are made available for the community to develop models. Those are benchmarks which are carefully shared. For the data, the good thing is that there are frameworks like Datasheets for Datasets [14] or the Dataset Nutrition Labels [15], depending on the type of your data.

I would suggest going through this, because it is a set of questions which makes you think about your data. For example, they prompt you to ask yourself if your data is revealing anything sensitive, if it can be misused, and how? They also help to, at the end of the paper, remind the community to use the presented dataset only for beneficial reasons and purposes, because almost any data could be somehow misused. So serve as a reminder, stating this explicitly. Answering these checklists might help the user to think if it is fine to share the data, or not.

GESIS: Our next question is about additional resources. You already gave a lot of suggestions about relevant conferences and papers. If someone is interested in learning more about this topic, are there some high-level tutorials or teaching materials or libraries or packages that you could recommend to them?

Sanja Šćepanović: So I am thinking about references for topics of health, digital health and so on. I think these links are helpful:

- The van der Schaar lab at University of Cambridge: https://www.vanderschaar-lab.com/prof-mihaela-van-der-schaar/
- Interpretable ML Lab of Cynthia Rudin at Duke University: https://users.cs.duke.edu/~cynthia/mediatalks.html
- Healthy ML Lab of Marzyeh Ghassemi at MIT: https://healthyml.org/

GESIS: **Last question, if there is anything, any type of research, artifact, package, an app, a method, or a whole research agenda that you would be free to follow, what would that be? Do you have anything coming to your mind?**

Sanja Šćepanović: Two things. One is going from connecting one dataset and some outcome to connecting multiple datasets to build a fuller picture, a very multimodal approach. Starting from my PhD, I have been dreaming to do that, but it has been proven to be a bit challenging. People start to do it, but it is not easy to have complementary data. This is really going to change a lot, because sometimes we might discover one hypothesis in one dataset and in another, complementary dataset another hypothesis, and this might start sparking among us the discussion on why.

> *If we can start to go beyond these simple associations and try to reconcile inconsistent research findings, this multimodality is one area.*

I can give you an example from another area I am studying; satellite indices for green areas. So you can derive green areas from satellite, it is called NDVI, a normalized difference vegetation index [16]. There is lots of research, association studies, that have linked green areas to better mental health, less diabetes prevalence, less obesity, and so on. But then, for all of those studies, you find at least a couple – and sometimes even more – contradicting studies. We have found that sometimes it is not enough just to look at how green the area is, but that you need to disentangle where this green area is, and who has access to it.

It is this multimodal approach: A green area in the backyards is not that useful, but green areas on the streets, which anybody can access, are much more useful. Moving beyond these simple associations by combining different dataset and modalities, and thereby trying to reconcile inconsistent research findings, this type of multimodality is one area that I am very excited about.

The second area is causality, which I already mentioned. A lot of our work is stifled in being applied because you will in the end get a question from policy makers if what you are finding is causal – because if it is not, we cannot apply it.[17]

GESIS: Thank you very much for this interview, Sanja!

References

1 Šćepanović, S., Obadic, I., Joglekar, S., Giustarini, L., Nattero, C., Quercia, D., & Zhu, X. (2024). MedSat: A Public Health Dataset for England Featuring Medical Prescriptions and Satellite Imagery. *Advances in Neural Information Processing Systems, 36*. https://openreview.net/pdf?id=CSJYz1Zovj [retrieved April 22, 2024]

2 Rausch, Z., & Haidt, J. (2023). The Teen Mental Illness Epidemic is International, Part 1: The Anglosphere. *After Babel.* https://www.afterbabel.com/p/international-mental-illness-part-one; Rausch, Z., & Haidt, J. (2023). The Teen Mental Illness Epidemic is International, Part 2: The Nordic Nations. *After Babel.* https://www.afterbabel.com/p/international-mental-illness-part-two; Rausch, Z., & Haidt, J. (2023). Suicide Rates Are up for Gen Z Across the Anglosphere, Especially for Girls. *After Babel.* https://www.afterbabel.com/p/anglo-teen-suicide [all retrieved April 22, 2024]

3 Global Mind Project, https://sapienlabs.org/global-mind-project/ [retrieved April 22, 2024]

4 Newson, J., Sukhoi, O., & Thiagarajan, T. (2023). *MHQ: Constructing an aggregate metric of mental wellbeing.* Sapien Labs. https://doi.org/10.31219/osf.io/d47q

5 Šćepanović, S., Martin-Lopez, E., Quercia, D., & Baykaner, K. (2020). Extracting medical entities from social media. In *Proceedings of the ACM conference on health, inference, and learning* (pp. 170-181). https://doi.org/10.1145/3368555.3384467

6 Gabrilovich, E., & Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research, 34*, 443-498. https://doi.org/10.1613/jair.2669

7 Stuchbury-Wass, J., Bondareva, E., Butkow, K. J., Šćepanović, S., Radivojevic, Z., & Mascolo, C. (2023). Heart Rate Extraction from Abdominal Audio Signals. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICASSP49357.2023.10096600

8 Bondareva, E., Constantinides, M., Eggleston, M. S., Jabłoński, I., Mascolo, C., Radivojevic, Z., & Šćepanović, S. (2022). Stress inference from abdominal sounds using machine learning. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 1985-1988). IEEE. https://doi.org/10.1109/EMBC48229.2022.9871165

9 Wang, Z. J., Zhong, C., Xin, R., Takagi, T., Chen, Z., Chau, D. H., ... & Seltzer, M. (2022). TimberTrek: Exploring and curating sparse decision trees with interactive visualization. In *2022 IEEE Visualization and Visual Analytics (VIS)* (pp. 60-64). IEEE. https://doi.org/10.1109/VIS54862.2022.00021

10 Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful. Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1-81. PMID: 34335110

11 https://www.covid-19-sounds.org/en/ [retrieved April 22, 2024]

12 Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science, 343*(6176), 1203-1205. https://www.science.org/doi/10.1126/science.1248506

13 https://dl.acm.org/conference/chi [retrieved April 22, 2024]

14 Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86-92. https://doi.org/10.1145/3458723

15 Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *arXiv*. https://doi.org/10.48550/arXiv.1805.03677; Chmielinski, K. S., Newman, S., Taylor, M., Joseph, J., Thomas, K., Yurkofsky, J., & Qiu, Y. C. (2022). The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligenc. *arXiv*. https://doi.org/10.48550/arXiv.2201.03954

16 https://gisgeography.com/ndvi-normalized-difference-vegetation-index/ [retrieved April 22, 2024]

All links in the text and the references were retrieved on April, 22, 2024.

Suggested citation

Šćepanović, S. (2024): *Expert Insights into Digital Behavioral Data for Health Analytics. An Interview with An Interview with Sanja Šćepanović* (GESIS Guides to Digital Behavioral Data, 19). Cologne: GESIS – Leibniz Institute for the Social Sciences.

Series editors

Danica Radovanović, Maria Zens, Katrin Weller, Claudia Wagner

Publisher

GESIS Leibniz-Institute for the Social Sciences, Cologne