Lind

Tue, 9/13 4:34PM • 23:32

**SUMMARY KEYWORDS**

data, measures, validation, framework, validity, studies, clear, case, methods, dictionaries, planned, review, measurement, understanding, hyper parameters, valid, standards, check, types, assess

00:00

Perfect. Okay, great. Okay, yeah, as I said before, I would then be very happy if you could give us a short

00:15

introduction into your understanding of measurement validity in the context of let's call it text, this data method is an overarching concept for all methods where you input textual data, and then you get some underlying score for constructs such as polarization sentiment.

00:39

Yeah, yeah, in my understanding, it's mainly focused at how the measurements that we obtain these Texas data methods, how well they fit the concept that we are searching for. So if it's only measurement validity, it's not really responding to inference to some population or some other truth. But it's it's really mainly talking about the fit between the measures and the theoretical concept that you define. And, and I think it's, especially in Texas data, it's often important that human understanding is has a key role in determining what's what's a good measure, and what's not a good measure? Yeah, I think that's,

01:43

these are the main points.

01:46

Yeah, thanks a lot. Okay, and how would you then describe the current state of validation practices within this field of Texas data? Research, from your perspective?

02:02

I think I can bring in maybe two perspectives. One is maybe from my experience with doing reviews in this field, and I would say that, from this experience, that validity is the thing so people mentioned it, often and to talk about it. It's not really the case, that is one framework where people operate in. And also that, I would say that even for, yeah, also for manual text analysis, it often happens that search strings are used, and they are not validated. And I would also say that that's also something where something is searched for a concept as defined, and then some automated dictionary search string type of thing is used to, to get the data or to get these types of measures. And I see that already is not common sense everywhere. So I'm not even at that stage, I would say that there's some, like, even in published articles, or things that get sent to represent there's one, one clear understanding from I think I can say a few things, also from the systematic literature review that colleagues of mine and I did. In this review, we assessed studies published between 2016 and 2020. And we coded them in respect to how well they do, how valid they argued for the data for the input that was provided for the aggregate route. How was was the process that validated and also was the output validated? And from that, I also coded a couple of studies there. So that's really published articles now, not only those that get sent to review, and yeah, it's, it's, yeah, no standards, really. But I think the most thing that's happening is that people do outputs validation so that they compare their measures of human coded gold standard, or compared with other variables. But overall, we find that it's only the case for 40% of the studies. But it's just includes manual coded studies as well. So not only computation, but it's similar and also it's also 30% of the computational studies to do this type of validation And and yeah, it's a bit dissimilar picture for other types of validating different steps in the process five to 40% in our systematic neutral view, use these validation strategies. I think people try and mentioned things, but it's not the case that everyone does it already on many stages, I would say.

05:33

How would you describe those differences in meditation practices between different groups of methods?

05:43

You mean, yeah. I

05:44

mean, groups as supervised or unsupervised methods in this case, are rule based. I mean, I think there's this this this trial key of Rubis, for example, dictionaries, unsupervised, unsupervised methods.

06:04

I would say

06:06

if people use dictionaries, and yeah, supervised classification based methods, it's at least easier, I guess, also to this output validation, because you're the standards a bit clearer. So that you're that people assess recall and precision. I think that's a bit standard. What I think is not very clear yet is for supervised machine learning. If it's okay to just split in training and test sets, and then just the test is already the validation. Or if you use another, like another held out data set for additional validation or some other type of data where you assess convergent validity. I think it's not clear fully clear yet. And I read different approaches to that. How to validate that and yeah, for for topic modeling, it's a whole different story, I would say. I mean, there's also papers out there and how to meditate on topic models or unsupervised methods. But yeah, I think they are. Yeah, it's, I think, understanding is probably that's, yeah, even more difficult to that. And yet, there are many variables in between many hyper parameters to select from. And I think it's not standard that everyone does everything that's proposed in validation frameworks for topic modeling, but people do maybe two or three things that are proposed and hope to go Yeah, or that's what they can do in the given time and with the resources that they have.

07:59

Yeah, that's that's a good argument, because this directly points to to my next question of, I mean, one goal of those interviews is to get a perspective, not just from reading the content of the paper, which is then ultimately published, but to get an understanding of the research process and the considerations when in doing research. And so this brings me to my question, do you think that there's some discrepancy between reported validation practices and unreported practices in the research process? Could you think of any steps validation steps, which, then, I mean, don't make it into the into the report or in the paper?

08:47

Death? for that question. I can mainly speak from my own experience, working in bigger projects. And then one example I developed also automated measures for I think, 12 topics or so and what's the plan, but seems to not really work out even on a manual stage, inter coder reliability, so we just dropped that. But it's not like, yeah, thinking back now, we could have reported maybe in some appendix at least, but we didn't. So we just concentrated on what we thought works, or what we could give some evidence that it works. And yeah, what didn't work was not part of the analysis anymore.

09:37

And I think Yeah, but what I observed for my own projects is that

09:46

reviewers asked for, ask more than maybe it was previously the case for yes, some checks and to also report What happens if you? Yeah, which was another set of hyper parameters, for example? And that's something that goes now to the appendix. I think that's, that's happening. But still not everywhere. And yeah,

10:15

I think some something gets some things gets

10:19

reported, but not everything. Yeah.

10:25

Yeah. Great. I mean, again, this is, I think, a good bridge to, to the audit log section here on how to improve the validation practices. Do you have I mean, you talked about the review start to acknowledge the the relevance or the importance of, in this case, robustness checks. And they do have you come across any other suggestions to further improve the validation of techspace measures? Measures are do it, could you think of any others?

11:04

I think what's really useful, and that's why I said, Congratulations for your project to kind of assess the status quo. What's what people do and what don't do yet to come to? Yes. To clear guidelines, what could be standards in the field that helps reviewers that helps offers entry readers? I think. And that's one thing and the other that the other thing that I touched a bit on was, is resources, or just validation? Things are extremely costly. And you Yeah, I think it's quite good. It's something that you need to plan ahead, like beforehand, and not just one validation type of validation test. But yeah, several ones. And many require also, yeah, the recruiting of manual quarters. And I think that needs to be planned. And that needs to be planned. Also, when, when people apply for grants. Because in my experience, it's not often where switch budgets are planned or planned for, but I think that needs to change.

12:37

So you mean by costly, it's mostly the mean financially costly for employing people to hand annotate data? Or was it the thing you were targeting at all other other things, which might be costly when it comes to computational power, for example? Or, you know, time time consuming?

13:02

Could because it depends on the models that you're running and the amount of data to store somewhere? I could, of course, also require financial resources that have to be planned beforehand. Yeah, just generally, the idea that proposing some automated study to funding agency and not planning costs to do that, that's Yeah, I think that can change.

13:37

Yes, there's one thing which I've like also wondered about, and I think that's, those are those two standards and frameworks, for example, on what things to test when you hand in a study using Texas data? And I also observed that there is no really like, then a gold standard. I mean, it's, it goes into the right time for that there is no commonly accepted framework. Do you have any idea what is the case? I mean, why there is no single, like dominant framework where people can like orient themselves on?

14:15

My personal opinion, I think we're not there yet. I think you're getting formulated at the moment. I think you just said gold standard. I think this is around for a couple of years now. And I think this is becoming more and more one thing where people can agree on it. That's an important thing to do. Because some benchmarks to compare measures to, but I think for, for example, pre processing steps, I think it's quite accepted knowledge that you need to contrast different types of pre processing and but it's not really that it's common sense that something doesn't get accepted. It's doesn't if it's not conducted or so it's not like a clear criteria to look for or how to do this. And yes, I'm not sure if if so easy to do that, though. But at least two guidelines could at least include your sections to prove that what I said earlier, like Data Selection, is justified even better, demonstrated that it's useful, backed up with some expert surveys, I think. Yeah, that would be helpful is based on reviews, as you do now. And as our team does formulate such guidelines that they can be brought in general, but still useful for reviewing. Yeah, make it easier also to know what to include and what to report.

16:03

Yeah, that's, I mean, this is not exactly targeted towards this interview. But that's what I'm thinking about my was initially blended this is my second paper, some kind of development framework building up on the review. And, yeah, I think it's just, I'm still in the process of understanding the level of like, complexities and different levels of analysis, when studying text data, which could potentially interfere with the with the, let's say, the validity of the framework. So that framework provides you with categories which are separate, like the so that you can separate it from each other branch off, and that it can be applied for all kinds of research. endeavors in the in the field of Texas data? So? Yeah, I'm, I'm really interested to have a look at your framework, if you are going to develop one as well. Because I think it's very much needed.

17:17

In some, yes, it's very, I think one one step is to kind of unfold all kinds of places where things can go wrong. Can can be checked, but then next step is like to kind of assess what's most important, or is there something like the most important and for example, is it open question like, if we check, output validity, so if we check outcomes with some measure, ideally, several types of measures, like for your example, if you want to use it for other types of analysis, too, so maybe check it again, against other types of variables collected for similar, similar purpose. But then like, if this goes wrong, then you kind of have to check pre processing only and not like, beforehand. And then maybe you're screwed anyway. But whatever they say is like, yeah, once all these smaller things are like, outlined, it's maybe easier when we have these frameworks to evaluate where to look first, and where to next, for example, to find arrows and to prove the measurement.

18:39

Yeah. There's, I mean, we're, we're getting to the end of this interview. But there's one really interesting remark of another interviewed person who says, I mean, one critical thing is that in the classical, for example, when you run your analysis is something that, for example, in hypothesis, hypothesis testing, there's no effect, then that's a finding. But if you run a Texas data analysis, and you're not measuring, like the validity of your measures bad, then there's no finding. So that that gives it a problem that you will need to derive a valid measure. And if you're not and then sometimes it might even be able to that it's not possible, for example, if you have highly high amounts of noise in your data, for example, Parliament parliamentary debates. And this is this is a challenge. And I've also thinking about about how to include this into some kind of framework. So the some kind of sequence of steps that if you are not able to derive a valid measure, then you can maybe try it with other things, but then to ultimately make the cut and say okay, this is not working. So we have to turn to hand annotation for example, because I think like in the lot of studies which have low validity, they are, at least and tweaking some some settings. So that like, it's like a pee hacking and then in some cases, they claim that they have a valid measure. Because they also have to pressure obviously if I mean, if they've worked quite some time on the, on the on the papers, or the projects to collect data, and if they're not able to measure something out of the data, then this this, at least, I think it's an incentive to, to just, yeah, go on, without being totally, totally sure about the validity of the measures.

20:45

This speaks a bit I guess, as to overfitting, I guess, I mean, I have most experience with very simple automated measures, like just dictionaries. And, of course I can, if I improve and improve and improve the list of words than I, then I can get to what I coded manually and I can kind of map to that. But what I would for this concrete case, always do and proposes to have another set of data where you test your test your automated tour can at least help a bit. But yeah, no, it's Yeah, I agree. It's looking like knowing what maybe just the end of it like deciding okay, this is the end of my trial and I have to start and like approach it differently is QUESTION

21:49

Okay. Great. And last question, is there anything else you would like to add to this whole discussion? Is that something which you might have might want to add?

22:05

Yeah, I think to to underline that not only measurement validity should be in the focus, but also yeah, having valid measures to do inferential statistics, I think that is an important thing to look at and yeah, because I did mainly multilingual aspects of course, I would also like to mention that and through Yeah, always, also consider what's the relationship between the concept and the context so that also part of the validation plan to see how bound is my definition to the concept country case that I study? How does the material fit to that? How do I train my coders if I work with manual coding, validation and if I have several languages, questions about equivalence and comparability should part of this validation discussion? Yeah.

23:27

Okay, great. Then I will stop the