Chan

Tue, 9/13 4:48PM • 29:38

**SUMMARY KEYWORDS**

validity, methods, paper, validation, measurement, measure, dictionary, topics, unsupervised, criteria, hypothesis, framework, question, model, data, people, based, research, talking, supervised machine learning

00:01

All right. Perfect. So, as I said before, and I mean, as you say, It's okay. Yeah, we're gonna do is it's a systematic review paper. And in this first paper, I just want to, like evaluate different, like screening literature in the field of political communication. And then to see, like, what are typical validation practices, also, depending on the method applied. And in a second paper, I don't want to build up on that, to maybe think about how to maybe build a framework or how to how to provide guidance in validating text based methods. And that's part of my PhD. I've just started in February. So that's the first part. Yeah. So maybe, maybe I would start with a question. What is measurement

01:01

validity? To you? In your research? Yeah, accurate for me, but they did they did it he is the correlate of findings that can be regarded as a very scientific effort. So this is actually the definition of validity, I believe in, in cube control. Yeah, but usually people will just saying that didn t is whether or not your measurement is actually measuring the truth. Yeah, so the deviation from rigidity, meaning that your measurement is deviating from the branches? Yeah. And but if you're talking about tax analysis, right. The variability in in tax analysis is a little bit different, I would say, because it is still measuring something, but usually, we don't know, we actually don't know the ground truth. And that's why the definition of validity is a little bit different. I believe that there should be three experts, at least I think most most of the people who are looking into what they call the statistical validity for example, like topic model, they calculate perplexity for example, these kinds of measurement is actually okay about as that historical statistical coverage of of the methods Yeah, whether or not that topics model can generate topics that are consistent progressive to basically measuring consistency, right. And that that is an evidence of validity. But this is not the only part of validity. There is another part of identity, which is called predictive validity, I guess. Meaning that the measurement that you create it, you can predict the world, the things going on in the world some criteria. No, this is this this criterion. Criterion validity is the first one which has semetic validity. Yeah, but predictive validity is talking about, for example, like, the sense humans can retake votes, for example.

04:10

Okay, so in a normal, logical or like, Yeah, I think the Atkins paper called it what was it called? Think construct validity and slash this.

04:28

Predictive, predictive validity, but yeah, usually we chose to do that in search questions, stuff like that, but I think what we actually care about symmetric validity, meaning that the measurement itself is measuring something meaningful, symmetrically making sense. Yeah. And for that, We need something more sample like criterion tests, meaning human coding or whatever, and then predict using using your your machine learning methods or text analysis methods. What a human understanding of meaning. And yeah, usually people do the first one, the second one, but

05:25

okay. Yeah. Okay. And maybe one question for the semantic validity because Mario's, as that's also talked about semantic validity. And I did some research and I couldn't really find this this term in the literature that much. And so do you think there's like some gap of like thinking about validity for taxpayers measured, measures this in this direction?

05:53

I think so. Yeah, I think they sent me two or three paper using this term, take a look at the module. And also Grima, and also one one king paper using this term. But most of the people with brand, they don't care about Radiata. T. They, they lump it all together? Yeah. And Emperor for the boundaries are that, yeah. So that they just calculate Professor T. And that's all. Yeah,

06:26

I want you to think that, like, how would you describe those differences in regard to different groups of methods, methods? So for example, supervised unsupervised goobers methods? Do you see that like any, like, significant differences between those groups of methods, when it comes to validating the method?

06:45

Yeah, I would say that a supervised machine learning is, is easier because it's based on some coded human coded data. And that's why it's predicting something that humans would consider as meanings in the in the tags. And that's why the semantic validity for supervised machine learning is somehow like, built in. Yeah. And then, but for the for the for the pool base method, for example, like dictionary members or unsupervised methods, that kind of symmetric relativity is usually absent. And most of the time, you don't measure. And it is difficult to do that as well. Because if you need to comb some data, then why bother doing unsupervised learning? Or dictionary based methods? Yeah. And it's somehow defeat the purpose of unsupervised learning. But yeah, with the core of identity, then Pete some people, somebody shows five to do that. But yeah, there there is some tips on how to do that. But yeah, we maybe we can talk about how to do that. Yeah.

08:18

Okay. And one thing we were also interested in, like, how would you describe the gap between, like, report like, venue and research process? So have like a certain method and you want to validate it? And is there like a gap of things you report and you'll find the paper, so some kind of checks? NSA also, like validation steps, which you don't report, like some decided there's a gap, and could this be problematic?

08:51

I will say that it's not something that is not important, but not done. Yeah, not done first and not done. And therefore, it's not important. And, and some people will do some some research or maybe do some, like sanity check. For example, like they would look at few cases. But they look at it in the presence of the unsupervised themselves. Yeah, not, not like creating a gold standard and then do the validation. Which I don't think is good. Because you already know the what the machine tells you. And then yeah, and then you do the coding. You may be biased by the results from the machine. Yeah, And yeah, finding patterns there. And then yeah, if it doesn't work, then do the validation. Again, code is always some more data again, something like that. Yeah, which is not a good thing. So, in my opinion, if somebody we want to validate those unsupervised methods, it should be. Okay, if they meet the gold standard, they should be done before the analysis. But for some methods is difficult, for example, like topics model? Yeah, it's difficult to do that. Unless you really want to do that. And when, for example, you code the topics first, and then do the topics model. But that would be I mean, why doing that? If you can do that, then why don't you for example, like create dictionaries or or or a semi supervised methods to find out those topics rather than make it purely unsupervised? And and most of the so called validation methods for topics model. I actually assume that the topics models already trained for some like no human in the new methods that you will find in long for example, of them assume that the topics model? Yeah.

11:40

I mean, I was watching my master thesis also about different ways to measure polarization. Yeah. And maybe that's the last question for this, like current dangers of validation and Texas data methods, but like, what do you think are the main barriers that this, like the main reasons that there's this this crisis? So for me, one thing I was always thinking about was about the selectivity of just picking the validity dimensions, which somehow support your argument? Yeah. Because for example, I was, I was conducting research, and then I wanted to compare it with like, survey data. Yeah. And I had the opportunity to compare with different survey data. Yeah. So obviously, like, for paper, like for a thesis, then I choose to choose the survey data, which somehow fitted my hypothesis. My expectations?

12:34

Yeah. So I think the major reason is that those methods usually have a have a have a methods paper. And then that methods paper usually saying that, okay, my dictionary, we were we well, with my with the data using that paper. And then researcher, we will then take the least resistant path, say that paper saying that, okay, this paper shown that this dictionary has some kind of rigidity, and that's why it can be directly applied to my, to my, to my research questions, to my research contracts. And, but methods, usually, I mean, the text analysis methods usually don't work like that. Because those methods dictionary, for example, but in particular, depends on the context. And that's why we still need to do the validations. Even with with evidence of validity for another context. Yeah. And, and this kind of context dependency is usually that flow. And then if there is some evidence that okay, this dictionary work in X context, and there must be, there must be a lot of paper also using the dictionary in the X context. And then we showed you, okay, like you said, Shall we packing the paper? They're saying that, okay, this state, what is one working X context, and my context is x. So that is work. Yeah, but your x is maybe different to my x sub toy that we don't know. For example, that x may be broad enough to fly. journalistic article they use but if you're talking about the use, there are different kinds of news. economical use, political use And those dictionaries are those are basically methods. Yeah. Again, it's context dependent. Yeah. But the context or the boundary of those contexts is usually not as concrete as we think. Yeah.

15:24

Okay. All right, then then maybe training a little bit on how to improve this problems. Yeah. I mean, I was wondering, why do you think there's some kind of lack of, like, coherent framework or like some kind of checks. For example, if you submit a scale to gauges of example, there's like, a distinct list of checks, which you have to fulfill, to show that your measurement instrument has some kind of counter validity, factorial validity, convergent validity. What do you think, are like to think that and why do you think that there's, like this lack of, like, general guidance or like, checks, which every Texas data method has to every Texas data study has to has to show?

16:17

Yeah, I think there is. Okay, lack of a coherent framework. I mean, a lot of people suggest different framework, but there is no. Okay, I will say that there are many frameworks, but there is no one that is very explicit about what are the requirements and how to do that, for example, like, if I want to validate a text, for example, we don't know how to calculate the required sample size, for validity for validity check. We don't have criteria of what can be regarded as useful. For example, if I found that okay, there is 0.3% correlation between the dictionary measure and the human coded measure. excusable? Yeah, I was wondering if, yeah, okay. You may, you may have a super large sample size, and then the P values may be fairly small, but the simulations useful. And then, yeah, there's a lot of these kinds of implicit rules. I mean, you can argue that, okay, 2.3 is useful, but how you saw is used? Yeah. And then there's this neck off this kind of framework, also the criteria and then there's also not enough guidance or papers on what to do next. I mean, if if my scale doesn't work then what to do? Yeah. Our paper that our there is one paper bias, which we try to think about how to what you do after dictionary based methods doesn't work. Yeah, but yeah, and then the next step is machine learning things like that. But But still, it is not a framework, it is more like a study. So, I will say that we need that kind of framework. And even with supervised machine learning, for example, after your your your training a supervised machine learning classifier, and then you find it the f1 is SEO concepts, if they useful Okay, based on their computer science criteria is kind of useful, but if you use it in a regression what will happen? Actually, that is there is some paper talking about what happened but most people will just regard it as okay, it's as good as the human code. Yeah, but yeah, we know that there will be something like correlation dilution. And stuff like that going on that will make the visa given your your your machine learning algorithm hat f1 equal to 0.6 or 0.8. Even still reduce the power of your regression. So, in most of the time we do regressions. So yeah, we, we don't know how to do it. First thing. Second thing is we don't know how the consequences and also what to do after you fail or what to do after it is assess.

20:25

Yeah. Okay. Great. And then last last question also going the same direction. Do you have like, any ideas on how to overcome the struggle? Like any concrete steps? I mean, you've mentioned some now, but apart from them,

20:42

yeah, I think we need a effort and base the framework on how to do this kind of validation by evidence base, I mean, we need to have some kind of like MCMC, MC, Monte Carlo simulation on what will happen if we follow these kind of validation steps? What will be the consequences of, of the end results? And then, if we do some modifications, what we will improve the results improve at least the simulated themselves? And then, and then we can devise a evidence based framework, for example, like, how large is the sample size out? What are the criteria and stuff like that? Do we need to do adjustments based on the results from the, from the validations, etc, etc. Yeah. And I know that there's some people doing that. And I'm somehow also in well involved in that. And hopefully, we will have something useful in the next few years. Talking about the simulations, and then even with the simulation, we need to have some best practices, articles, and also software to enable these kinds of these kinds of practices. And I think there's one, there's also one very important thing is that there will be a lot of these kinds of best practice articles out there, but we must have some people to enforce those practices. Right. They are still a lot of paper Partridge without any validations. But yeah, it is because of the lack of enforcement. I think, for example, like a journal or a conference, should actually enforce these best practices, when they ask their reviewers to to, to review the articles for that journal or for for that particular conference.

23:30

And I mean, that would also imply to have some basic understanding of what has to be done. I mean, not just some, I mean, you could you could say you have to validate your method, but then that basically could be anything or you could say, there are some like certain steps, which you have to take for validation. See if the theory and methods somehow fits. If your measurement correlates with some other measure or some external criterion, yeah. Yeah. And apply this universal framework. So

24:09

yeah, that would be great. Yeah. Yeah, I yeah, I try that. Like, topics model I try to manipulate or kind of where we oppose that. Could that could change your results and then try to pod universe.

24:34

It's an interesting I fractured the paper, about a topic model validation.

24:40

Yeah. Yeah.

24:44

Yeah, great. Last question. Is there anything else you would like to add to this whole universe of topics something which is burning under your fingers, which you went to?

24:55

Yeah. I am. Now, I'm actually a little bit pessimistic about what to do next. I mean, the what to do after something fail. And I will say that it is severe, severely underdeveloped? There is a lot of paper, there's a lot of research, talking about the rigidity problem. But what to do next? Yeah, and actually there are lots of methods are available in the statistics, literature, for example, talking about how to do those kinds of adjustments. And I will say that these are the low hanging fruit that we should consider, rather than trying to make something not very that

26:04

one of the things about pre registration, is it, I've got a feeling that it's not as common as in other disciplines, isn't it?

26:15

I mean, for political science, and also communications, there's no requirement on pre registration. But pre registration also. Okay, I support to register. But, but yeah, then we will need to think about what would happen if I pre register something. And then the validation check fail?

26:46

Because it's not really supporting the, I guess that targeted towards the hypothesis, but what's the measurement? Is that

26:52

Yeah, because, yeah, usually we pre register the hypothesis. And we also pre register how we are going to operationalize the construct and those hypothesis. But yeah, once again, what if my operationalization fail? wants to do? And? Yeah, no, one? No. And actually, the standard answer to this question is, you can deviate it from the hypothesis that the pre register hypothesis, but you didn't, you didn't you didn't solve the

27:36

problem? Yeah. Yeah. I mean, it makes sense to start with unsupervised methods, because I think that there's like this very, like, established pipeline that you hand code, some some sample, and then you try to do like, some kind of language model? Yeah. All right. Yeah. That's a good I haven't thought about this preregistration hypothesis or progress in a motivation problem?

28:08

Yeah, we actually have a similar experience, and which is we pre register some hypothesis, and then we code a lot of data frame, some supervised machine learning models. And we found that Okay, two of those model doesn't work. The one

28:29

and multiview.

28:31

And we don't know and the solution to this is, we deviated from the pre register hypothesis, register protocol. And then we just analyze those coded data. And those coded data actually randomly selected from a whole corpus of data and then do the analysis. We can reject the null hypothesis but but covering up always for we why and then it basically saying that the supervised machine learning part is not required. Yeah. Yeah. Yeah. Why bother? Yeah, there's actually a paper by form and Taylor about that fairings relation?

29:33

Great. Think that I can Yeah, roughly