Gespräch Stier

Tue, 9/13 4:47PM • 19:03

**SUMMARY KEYWORDS**

data, validity, methods, models, measurement, field, texas, report, research, result, sentiment, f1, parameters, hand, descriptive, link, exploratory, benchmark, algorithm, paper

00:00

informed consent. amazing to me.

00:06

Yeah, we can, we can. All right. And so yeah, thanks a lot. So as to said, What is your understanding of measurement validity?

00:19

Very good question in the context of Texas data, I would have to have a look at crema, Stuart, again, to differentiate these dimensions. I mean, they are the standard 10 dimensions of validity. Obviously, that should also apply to Texas data. I think one of the most important things is that in terms of just this kind of goes into the direction of face validity, the measurement can establish the link between the relevant theoretical constructs. And the empirical patterns in the data and would say that is not trivial. So, the early days of the field, myself included, on my own research included topic models were the go to tool, because they allowed you to quickly find patterns in the data, meaningful patterns in the data, even topics that latent concepts that represent real world conversations. So that was super exciting. On the other hand, the papers that were, were or still are applying topic models, most of the time pretty exploratory because there's no direct link, or topic models do not allow for direct link between theory and what you're actually looking for, and the results that measurement is producing. So I would say the most important thing is that you choose a measure that allows you to make this link to theory or even better hypothesis. Pretty clear. And that regard and supervised methods, I think are ripe now to go to tool, if you want to really move the fields of service science, Texas data inquiry forward. So that was not the most systematic answer in terms of which validity dimensions or whatever are important, but rather, lessons learned. Yeah, answer from the last few years of exploratory inquiry using Texas data methods. Yeah,

02:31

I mean, if I want to have a textbook definition, then I can just open a textbook. So like, generally, how would you then briefly talked about this, describe the current field of research within Texas data,

02:45

super exciting. But on the other hand, so at least my impression, or the stuff that I'm seeing pretty good signs, in terms of the topics that are covered, pretty one sided also. So there are a couple of dimensions to that first, and the field of party politics, parliamentary speeches, etc, is, I think one of the major drivers of innovation in this field, at least from the political science perspective. But on the other hand, they miss out on other relevant text sources, oftentimes, even social media is just used for, let's look what politicians say. And that regard, I think there is more potential in terms of the substantive fields, where Texas data can play a role. So I think that needs to be topical, expand the beyond the usual suspects, at least from a political science point of view. And the second thing is that the methodological innovation is great everywhere, everyone is now using bird and transformers and whatnot. But on the other hand, I fear that we lose some of the face validity that I just mentioned, because methods are such black boxes. So in the end, that that is what Mary's tells me, no one knows or understands what bird is doing. But in the end, you have a result that looks decent, or that has a great f1 score, whatever. But I fear that by not understanding what is going on, we kind of weakened this link between theory and measurement in the end. So that is, I think, a broader trend that is clearly obvious that methods become ever more fancy. But also more intransparent.

04:35

Have you come across any practices to to target this problem by finding new ways of validating, for example, for example, like feature selection or stuff like data?

04:50

That actually goes beyond my expertise, to be honest. So my personal takes this data experience stopped, I would say one to two years ago. Oh, when Bert was still in its infancy, I will say, values, there's lots of great things. So he oftentimes looks at features, he can include different weights into the training process, making the process more representative of the training data, stuff like that. So I think there are lots of different ways how you can tweak these models and better understand or better fine tune what they are doing. In terms of validity, or evaluations of the output, I think more work needs to be done in that regard that we better understand where these models tend to fail and where they, they can really outperform, let's say, more traditional stuff like support vector machines, Naive Bayes, or whatever. And I think if you have a clear benchmark, let's say Naive Bayes, we know straightforward, we know what it's doing. And if these methods are a benchmark, where we understand what is going on, and then we look at individual dimensions where bird stuff like that improve on knife base, I think we can get a bit of a better understanding of what mechanisms inside Bert lead to improvements or do not lead to improvement, stuff like that. But again, I mean, the need to be clear here. I've never once applied a birch model. Now I have Mario's and Felix Schmidt, as an affiliate Smith, right.

06:31

Yeah. Mentioned that truly crazy

06:33

stuff. Okay, yeah. But I mean, we that is really something important. I think we need to better understand what drives certain performance gains and better look into these black boxes.

06:48

Do you think that there's potential besides those supervised methods, like speaking about, for example, like rule based or unsupervised methods?

07:03

Of course, there are potential but it always depends on the research interests. And I think if we just did to, I don't know put it put it bluntly, this exploratory let's run a topic model or run some supervised unsupervised method on a data set. This is what I would say CSS 1.0 so distinct from social science theory driven research endeavor. Let's see where we land in the end. But I think the way to go is really to fk research, questions, theory, maybe even hypothesis and for that, you you actually need to provide methods I think. Yeah, unsupervised methods should play a role. But again, it always depends on the research interest.

07:51

So when you think about like a typical classical research process, would you like for example, in your research, do you think of any practices which you do within the research process, maybe extra exploration or stuff like that, which might then not be reported as a final result in the publication? And how could you describe them? Also, in terms of measurement validity,

08:20

I guess we always do. So. If you have labels, on on data, topics, stand framing whatnot, sentiment, the first step should usually is to look at differences in vocabulary between these categories, to tf IDF, whatever exploratory stuff, the My take is usually in a try to be consistent that regard if there is no descriptive pattern, the project is that basically, so if I just bet on one or two significant stars, that are the outcome of a regression model, where lots of other staffers also include this input, that is, I think, too weak of a result, please find out that is I would say is should be the gold standard, if there are no descriptive differences between the groups or categories or whatever of interest. This should not be published. And therefore, I strive for also having descriptive results in papers or in the appendix or whatever. So that whatever we do descriptively is usually not definitely not everything is reported. But descriptive findings or differences should go hand in hand with what is the result of your 100 regression model?

09:42

Yeah. That's an interesting missing point. And I guess this already leads to like, like the question of what would you suggest to like, if any ideas on how to overcome those challenges associated with measurement validity? Tea. So you've mentioned descriptive statistics as some kind of being transparent and making making sure that your method actually measures something. Even if it's not maybe percent a construct of interest, but it's still like some systematic differences.

10:21

Which brings me brings me to one thing we have not talked about all parameters and optimizing parameters. An LDA topic model has several parameters, you need to choose nine to whatever as an input. Each and every one of these methods has some parameters where you have insane amounts of research or degrees of freedom and how we transparently report, test optimize across these multi this universe of different combinations of parameters. That is also I think, one one challenge, I also struggled with my research. So for instance, a recent project with Felix, we use an off the shelf sentiment, algorithm Raider, which we were very skeptical in the beginning. The project is on negativity in the Buenas tardes campaign last year. Just negativities is something that you can put on a continuous sentiment scales and we can keep we can link sentiment to the actual research interest. Sure, why not try sentiment dictionaries or go beyond sentiment dictionaries by using wader. And it works. The problem is now we also have to choose thresholds. So Vader Raider has predictions, but you need a binary variable is that sorry to validate it. We have lots of students systems, that project partners team who hand code data is their negativity also on a continuous scale towards certain objects like candidates. But for weight, Vader gives you a probability and therefore you again need to pick thresholds that can then be linked to Student Assistance codes. And for that, what we now did is we ran 1000s of models, choose the optimal parameter that maximizes the f1 score. And now, we have a nice result. But on the other end, it still feels like a bit like pee hacking. And the question is, in which context, is it? Okay, to just maximize your f1? And in which contexts, do you have to scrutinize these decisions more? I think in our context, I think it's fine. We use this we have TCB codes, and we use the best performing model to demonstrate and validate that wader measures the same thing like they do. If we would build a training data set from the V codes, I think that will be more problematic, because then you have some endogeneity if you tweak the algorithm, and link the algorithm then to even more heavy code. So if we would then say okay, code to under more of these posts, stuff like that. If we would have an iterative process, so to say in the training, live the algorithm, then it will become more problematic, because then you you fine tune or maybe even what's called what's the right word. algorithm that performs well, super well, on one day does that overfitting thanks, if you then if that leads to overfitting the external validity gets lost, that is fine. So that I'm sure you know, this guy, Sperling, New York paper where they show how the parameters that you choose before running a model makes can make a pretty large difference. I think they are the proper reporting standards. And what is considered as an good solid documentation of Texas data methods in the social sciences is still lacking, I would say maybe that's one of the outcomes of your project. Standards Guidelines.

14:14

Yeah, I mean, that's, that's that's the plan for paper two, but maybe speaking more generally about those frameworks or standards because I get the feeling that like already some some stuff out there's already some stuff out there.

14:34

Lots of stuff about completely lose track the field. As I said, I dropped out because of other stuff, writing grant proposal. writing emails management stuff, and then you cannot catch up. I mean, I'm so glad that I have money who's who's really at the forefront of things I would say? It's tough

14:56

to do. Like I was also We're hearing a lot about those cutoff values, etc. Do you think we need this as well for the tests of validation? So different dimensions of validation? For example? I mean, typically like comparing your estimates with an external criterion. I mean, if you're just supervised and you usually have your hand coded estimates, do you think that there should also be some kind of general framework on which validation steps to report because at the end, like, like, we're asking ourselves a little bit, and obviously, from my own experience, as well. I mean, if there's like 100 opportunities to validate a model, like from from different comparing it with different data, for example, then there is no need to report the things which are not working, but there's an incentive to just report things which are working. And for example, if you look at psycho psychometrics, if you validate a scale, for example, here, in this cases, my team, then there's like specific steps, which you have to show that your instrument fulfills its requirements. And I mean, there was one thing we were talking thinking about that this was completely lacking for, for Texas data,

16:16

super desirable to have that. But the scale of scale development, the dimensionality is really a lot less complex than Texas data methods. Were very self pre processing decisions, fine tuning choosing models. I mean, there are 5000, supervised tax analysis classification models. So do I have to testify that I use this instead of that? And do I always have to report a full range of results? hores, you try five different algorithms. There's so many decisions that you take during this research process, which usually also takes you two years and not just one, because the data is so messy, and you have to clean it, and you have to throw out words, throwing out words. I mean, this is a radical decision that we just removing stop words we bought in a paper. Yeah, it's been done before. Yeah, I mean, it would be nice to have such benchmarks. And if cases can produce that, that would be super amazing. I'm sure there are papers doing exactly that. In some field, we do not even know about this problem that so many people are right now working in this field. The paper is out there somewhere. It's just hard to keep track of what was one thing you also asked. Yeah, you need that.

17:38

Okay. Yeah, I think it was some kind of lists of validation aspects or steps of validation, which you have to show that you at least did it and then report on if you did it or not. All right. I mean, is there coming to the end? Is there is there anything else you would want to add to this whole measurement validity, Texas data? Problematic problem.

18:03

There's so much potential but also so much going on. I mean, if you're doing a PhD now exclusively in this field

18:13

must be super hard, super hard to find your niche. Right? What can I contribute to this insane thing? It's been tough. Good luck.

18:25

Yeah, I mean, I think this measurement validity is already you have a

18:28

meta perspective. Yes, I mean, exactly that just take a bird's eye view. Look at it from a meta perspective. Just try to make sense of all of that, but not try to improve f1 by 0.02 or something like that, that that is what computer scientists are doing right? Benchmarks then we outperform this and that that is not what we should do. I think we should we need to order need to know what is useful in theory. Validity