Baden

Tue, 9/13 4:47PM • 40:58

**SUMMARY KEYWORDS**

validation, method, text, problem, measuring, validity, standards, classification, reported, assumption, f1, people, analysis, question, dictionary, pre processing, sense, unreported, models, develop

00:00

Interview interview, we broadly be on three topics. So one, just very briefly definition of validity, then you're like the current status of validation. So how would you describe the current state of validation? And then maybe some some possible outlooks on how to improve validation practices in Texas data research? So yeah, the first question would then be without wanting a textbook definition, but more like a personal definition from your, what is your understanding of measurement validity? In the context of Texas data?

00:41

Well, I think, well, the core of validity will always have to be that there is a close correspondence between the conceptual understanding of what we are interested in measuring, or believe we are measuring, right, but sometimes we know we are not measuring the thing we're really interested in. So if we decide on a proxy, that's indicates something else we're interested in, then there should be correspondence between the meaning of that proxy, that's the thing that we think we are measuring. And the thing that we're actually measuring, and in both sides are non trivial, right? Because the conceptual definition generally has the problem, that concept on text, the text doesn't collaborate, right? We're approaching the text with a set of questions that the text most likely never had. So there is a need to anticipate in the conceptual definition already, that the text does not necessarily want to answer the question that we have, which may lead you to proxies, which may lead you to rules about how you deal with vague, insufficient, elliptic, lalala kind of data. So that side, the conceptual side needs to anticipated text as text, and it will not cooperate. And the on the measurement side, it's, it's non trivial. Because if we argue that there should be a correspondence between what we believe we are measuring and what we are actually measuring, then the problem arises that what we're actually measuring is not clear. Either basically, that kind of text, the stuff that is gets an annotation in our text, may or may not be a inter subjectively agreeable case of the thing that we have an interest in like just to raise one thing. polysemy is a huge issue in text, right. And that means that just because one coder has high confidence, that this sentence is a case of this construct, that does not necessarily mean that this is objectively so. Right. So like there is a certain gap between the manifest content of the text. And the interpreted content, as read by a coder or as read by a machine, or machine doesn't interpret in a narrow sense. But a machine draws conclusions about patterns. That permits for some ambiguity. So there's this kind of like this. It's a bit like kind of like a correspondence between two moving targets, right, like two or two fuzzy things. And that makes it difficult. But I think that's the core of it. And then I would furthermore say that for a valid Texas data analysis, it would be necessary that one can define on the conceptual level and demonstrate on a textual level the boundaries of the coding, and that is something that I think is more important than of giving credit. Because any idiot can hit a core case of a category, right? Like if you want to code whether text talks about freedom, and it says freedom. Any idiot can do this. But the boundaries are the problems and like, I would include in a definition of validity that it stops calling when it should stop calling.

04:30

Okay, there was a lot of input. Yeah, but it's interesting that this, this simple concept is so multifaceted, because I'm like, You're the fifth or sixth interview interview. And all of them got different definitions. Apart from the obvious like measuring what it is it is intended to measure. Yeah, but turning to the validation practices, how would you then describe the current state of validation practices within the field? Oh,

05:02

I wouldn't make a few differences, I don't think that there is one set of practices, there are a few cultures of validation. And they're good at different things. There is one thing that looks a little bit like a salad. But it's very useful. And that is this kind of very sincere and chance to probe the coding the classification in as many ways as possible to gain confidence, which often were Wilson's slightly unsystematic like combinations of convergent validity and, you know, the predictive validity and like different kinds of things. But as useful. And it is useful, not necessarily because it achieves its gold particularly well. But it is useful because it thinks about how one could do this for a particular case. So these are people who, and you see this every now and then, like they basically they have this measure that they propose that could be whatever a dictionary could be a supervised classifier. And then they kind of pull one thing after another out of their head in order to try and kind of make data criticize the algorithm that they were using, they kind of critique the dictionary of the kind of the input features and so on, showed that there are certain kinds of cases that are over coded, they do error analysis. That's cool. It's unsystematic. But it's cool, because it shows what different approaches show you different things, then there is the canonical approach. And that's basically f1. Right, kind of having a few hard coded things, comparing it to what a computer does, and believing and I think incorrectly believing that f1 in excess of point seven or whatever it is good. And I'll immediately say why I think that's not correct. I think this assumption that point 7.8, or something is good comes from statistics. And in statistics, the assumption is normally that if we get point seven, right, and we have point three random error, then the random error is white noise and white noise disappears in the analysis. Or the problem is that Tada analysis or like generally anything computational doesn't do random. Whatever mistake that makes this systematic. The question is whether it is a systematic error that correlates with the thing we try to measure, or whether it doesn't. But it is completely conceivable that you have an analysis that gets an F one of point eight, but completely misconstrues an entire set of the debate, and you would never know. And so that's basically I have a huge skepticism towards this f1 strategy. It is a good start, right? Like basically, if you can't even get you have one up. That's not good. But I think that's far from sufficient. But it's kind of the cheap and easy solution that most people use, because they confuse statistics with with algorithmic classification. And and the good cases of these do some error analysis, because that that does something, right. If you have point eight, right, and you have an error analysis that informs you what kind of misclassifications you get, then you might get a handle on the biases that you have. So that's kind of the second kind of big basket of validation strategies. Then there is like a, like, I would say, maybe a methodologists approach, which doesn't actually start from the achieved classification, but it starts from the algorithmic design, right, like, kind of thinking about it, what kind of feature design should be working well, what kind of hyper parameters should I be setting by like approaching validation? Not so much as a parametric axis or not solely? I mean, it's an add on stage, like people do this normally do some of the other things too. But like, not thinking of validation so much in terms of I apply a classification and I'll check whether it was valid, but more in terms of checking the tools that I'm applying, and trying to figure out whether for a particular given task C, A bigram, based representation with a recurrent neural network classifier makes sense. Well, Whether that's plausible,

10:03

from what we know about language from what we know about the constructs that we are using, whether it's plausible that this kind of algorithm should be capable of detecting this kind of pattern that is unfortunately rare. And it most likely fails at some point. Because at the point where, like for dictionaries, you can do this for NLP, you can do this to some extent. For some unsupervised models, you can do this. I mean, like, for instance, you, you can build arguments for why topic models of this or that configuration should be a valid representation of the topical structure in the data. It all falls down, of course, in the point where like, the assumptions that lie and lethargy and all these other people who develop them don't make any sense. But well, you know, there's something. But for supervisors, it falls down, because we just don't really know, there is no systematic knowledge that I am aware of anywhere, either in computational computational linguistics and social sciences, about the comparative advantages of different classifiers other than a few rule of thumb experience values, that we like, for instance, to get a good or economically sketch, trying to work with a Naive Bayes classifier, because they almost always suck. But but but like, you know, like, if I want to make this this gorgeous study by Hearst, and all that kind of try and classify ideology in parliamentary discourse and end up classifying government participation, right, excellent, accidentally, which is really smart, because they noticed, right, and it's the difference from other people who just apply the thing and just assume that they're good. But but the problem is that at some point, like you can, like we do have some capacity of deductively, checking whether our training sets are set up reasonably, we have some capacity to deductively, check whether our pre processing makes any sense, right? Like if I step away, or grandma, does that makes sense, given my research question, that we have zero knowledge about which classifiers we should be using? And that's the point where this all falls down. But I think that's essentially the Okay, oh, yeah, well, maybe you have to, if you want to, if you want to really be generous, you could say there is this kind of fourth class of validation approaches, which is looking at the results and feeling satisfied by them. So like, this kind of thing that I don't know, who forgot the name, but somebody called this the researchers degree of freedom, right? Like you run a few models, you look at what it does, and you pick the one that looks good. I wouldn't count that as validation, that's just seat of the pants bullshit, but but it's also a practice that is common. And in terms of commonality, I would say the one magic dominates then you have roughly equal size, putting anything like this, like, you know, kind of seriously concerns mode of validation, like trying to do everything and the looking at it and feeling good about it. And then you have a tiny marginal percentage of people who look at the algorithm.

13:41

I mean, you've mentioned a lot this researchers degree of freedom, which which have also come across a lot and and also the this divergence between reported validation practices and unreported validation practices. Can you maybe describe in your in your words within the research process? What do you think like challenges are just degrees of freedom? Yeah. Which which might be not reported in the final papers then or in the final project?

14:14

has multiple layers of that one of them is that like for some validation sensitive practices there is reporting that these steps have been taken pre processing is a classic thing I think most people's that pre process mentioned that they pre process that don't mention why they don't mention why these steps but they do mention that their pre process was some way that this thing is consequential. There's this gorgeous paper by Damien Sperling of shows that is really, really, really, really matters. And that one should really, really, really, really validate what's pre processing stages. But that is almost never reported. So like, in a way, that's one thing I would say that there is like a ton tendency to report steps as if there were technical and neutral. And ignore that they actually require validation. And then focus validation on the output. This is like the this is like a classic thing in this f1 style validation, right that basically, you assume that everything that you do is more or less neutral. In the end, you get like one score that shows that you will write, but you never test your assumption that everything is neutral. That's one thing, I think then there is, it depends a bit on the method, like for instance, dictionaries, I think do normally undergo more validation than was reported. Because validating dictionary is such a bloody nightmare. And it just takes like, there's so many small things that you can do. But, and like writing them up always feels a little bit stupid. And you can fill an entire page, which the reviewers will tell you to take out anyway. But you definitely did that. And that again, goes unreported. So I think there is more validation in the process of the tool in dictionaries than we normally see the same thing to some extent goes for NLP methods, where normally people test like, you know, like, most people I know who use like whatever pauses, tigers, stuff like that. They run them, they see whether what they do makes any sense. And they go through this tech pull some quakes and see whether they are correct, they pull some descriptors and see whether they make sense that almost always goes unreported select algorithm choice validation goes unreported normally. And then basically what's the you have this, like I would say like, in the more ambitious corner of the predictive validity, like the f1 stuff, people do report. That's, that's right, a number of models. They do report some coefficients that express some predictive validity for these. And that goes both for like sometimes you have to search for topic models are slightly less common that you give like the distinct Miss Woodleigh. But for supervised, it's pretty common that you have a few models reported. But then the error analysis and the qualitative validation if you want, but they do kind of checking that it's not only, like when you can get like a high f1 from agreeing on absence, most of the time of rare categories, right? Like if, if a category is absent 90% of the times and the classifier codes have not a single time, you have point nine accuracy. That's not very useful. So, so that kind of all the measures that actually go beyond the quantitative, they normally go unreported. And if they are reported, they're only reported for the model shows. Where one doesn't know, right, it's completely like I mean, I know from my own experience, we've done a few cases with topic models, where we actually did go quite deep into the alternative estimated models in order to check that they make sense and that the interpretations that we gave to the topics make any sense and stuff. But in the end, we only reported it for the model that we chose because you're not Regulus length. So I think that that's that's one issue that matters. Last thing I should mention is that validation goes not to last things. Validation goes unmentioned whenever somebody uses a tool that somebody else already developed.

19:09

So they just assume that if you know like somebody validated to look categories for something, so you just put them onto your own new data and you just assume that they will work on the status that's there. I suspect there's actually no validation takes place. But one doesn't know. And the other thing is, Texas data methods as auxilary methods almost never received validation. But basically if you primarily do something else, right like for instance, you have like a survey study and or you have like an experimental study and you have some auxilary use of Texas data that identifies the dominant frame that identifies lead or whatever right. Almost never validated

19:59

it I'm just just for time management practices. And we are a little bit short on time. Do you have like any preferences? Do you need to leave now? What I can fly for 10 minutes?

20:11

Are you? I have my next meeting is at 1110. Your time?

20:18

won't be that long. But great. Because like, there's one specific question which which was brought up in the in the interviews before? Maybe also having an a perspective on how to improve practices? There is this. Let me let me Yeah, I'm just thinking about when it makes sense. I start with this one. So the there's one point that when you apply a Texas data method, and you see that, like that it's basically not working at all often are like, there's there's this problem that other than, for example, in pre Reg, registering a study, where you say, Okay, there's like this is your study design. And then every finding is research funding. So if the if your hypothesis are supported, that's a finding if it's if they're not supported, that's also finding, but that I think it's more difficult for text for for the measures, because if your measure are not supported, then your whole research design basically is is weak. Yeah. Ideas on that? How to solve this problem?

21:35

How to solve it? Well, I think the solution for that is a longer process. I mean, I do agree that basically, if the reason why your expectations don't come up, is because your method sucks. That's still interesting, right? But you should not sell this as a no findings, because it's not a finding where the problem lies before finding. But I think the I don't know, like, I should just put a very quick insert in there. I'm not a huge fan of pre registering everything together. Because what it does it it kind of pushes data into this hypothesis testing. last stage research, right, like kind of like when basically all the thinking has been done. And now we just need to ascertain that our thinking was correct. And I think there's a lot of potential and other methods for CO regenerating, for hypothesis generating, that doesn't go well with preregistered stuff. And of insert, how to solve it. And I think that goes a little bit in the direction of the third block of questions that you mentioned, what we should be doing, I think, is that basically, we really, really need to think very seriously about Tara methods development, in terms of validity. Because what and that's kind of what we wrote in this three gaps paper, right? Like one of the things that really bothers me is that there were very few validity considerations that go into the method development. And as even with stuff that actually claims to be specialized, right, I mean, I say play unless it is Article doesn't even mention a single social scientist who knows what topicality is, and yet those people have the audacity of calling it a topic model. They don't know what a topic is. If you can't define what a topic is, how on earth can you develop a topic model? I mean, yeah, it has some plausible assumptions, I can see why. You can see where they're coming from, right. It's not absurd. But this is not how you do this, like if you want to, if you want to make a claim, and that goes for every method, right? That like you develop a tool that is capable of detecting a specific kind of pattern, semantic, pragmatic, whatever in text. You need to consult the knowledge about how that thing is expressed in texts, while developing the methods you need to and you need to reflect upon that. And you need to do this together with applications just kind of building an algorithm, running it on a text, then staring at whatever it happens to deliver. And just after, you know, like kind of after doing this workshop, digital wash off test for a half an hour, deciding that like this looks like a frame that's just not good enough. So I think that's the that's the that's the key twist that we can do. Because if we think method development, about validity, that means that the points in the method Make consequential assumptions will be explicated. So I can check them while designing while applying a method. That means that I have a vocabulary for discussing, like mismatches, like you know, like whatever you design this method that recognizes, like what you said like for instance like, you know ideological scaling political text if you while doing this, also talk about what is political texts? Like what do I mean by political texts? am I assuming like manifesto style text what is characteristic of manifester style text? Why does this assumption make sense here? Why does the other note make sense that it is easy to notice that the very same method might not work all that? Well, if you apply it, apply it to Twitter. And then the problem that somebody uses a method and suddenly finds out that the method has no traction on the desire construct, shouldn't really arise unless you're an idiot, which of course happens, right, but that's not my problem. So I can have all these problems that arise further downstream. And that's kind of the I think the foundational misconception about validation that we have in computational tools currently nested within a data at the end, what we need to validate in the beginning, we need to validate when we choose the method when we designed the pre processing when we when we develop the method, because if we do that, then we already have reason to know

26:34

what things they will, this method will catch validly and where the problems will lie. And we can choose much more intelligent error analysis strategies, we can choose much more intelligence post hoc tests in order to check where we are.

26:51

Yeah, yeah, I think most of the time, it's somehow implicitly mentioned that if you say, you apply a method, which is already validated before, then you you implicitly argue, okay, that's the correct selection of method for your specific task without being like, for the reasons you mentioned.

27:10

Sometimes this works, right? I mean, like, you do you do get variation, right? Like if you if you pull a few throw a dictionary developed for political manifestos on Twitter, you do get measurements. The question is, what are you missing? Because social media really doesn't talk like manifestos. So there is a theory likely situation that was a significant share. And what is more important, is systematic share of the stuff that you wish to measure that you're missing. That is the problem, right? Like the problem is not necessarily like a method that has been designed to do one very specific thing normally can do this very specific thing, as far as the assumptions that the method makes are met. And the moment is assumptions break down. Computers do exactly what we told them, right? Like if a computer, if you tell the computer to follow the dictionary, it will follow the dictionary it will ignore every misspelling every neoliberalism every slang every abbreviation every. It will just do this with perfect reliability. It will not tell you that it did that. Because we obviously did that. And if there is a part of your corpus that relies on these things, you will never know.

28:35

Are there any other suggestions? Apart from this? Not password validation, but thinking about concepts or like theory at the beginning of method development, which you would propose to improve the quality of validation practices?

28:55

Yeah, I think like, I mean, basically, one thing we really need to do is to create a space for justifying processing steps. Justified processing steps as part of validation, right? Like having a reason for why you remove these stockworks is part of validation. And we need to think of it as as part of validation, because it really affects the outcomes. If you do it wrong, you get invalid results. So it's clearly part of validation. And that's a cultural thing, right? That means that affects the journals the journals need to tolerate if you do this, that too does the reviewers need to demand that you do that? And you need to be enabled to do that. And it is almost the most difficult thing because as long as all we have is a little bit of like rule of thumb seat of pants experience values. There are no real theory about why which processing steps would have what kind of implications this is bloody hard to do, but that's where we need to go I think we're like every in every algorithm that you apply to the data from retrieval, through cleaning, through pre processing, through modeling, through classification, every step effects affects the outcome in systematic ways. There is no random error in this entire process. And that means that for every step, we need to have a justification for why we did it this way.

30:31

Maybe just a brief comment on that, did you know that I've come across this pretext our package, for example, which basically does, calculates all different outcomes for I think it was, it's about fish model for like, all kinds of options, like removing punctuation, etc. So I really liked this. The idea behind this, this package, and I also applied it in my thesis

30:58

is basically this kind of idea that that means Burlington used to demonstrate that really every variation that you do that really makes a consequence all the way downstream. And that's the starting point or notice that this matters. And the next step, then is to understand how it matters. And why. Because if we had some points at the place where we know that, for instance, for a certain kind of measurements, for a certain like, you know, take something non trivial, right, like if you have programmatic analysis, you're interested in like interactive behavior, expressed subtext. And if we know that for this kind of thing, punctuation is essential, we really need to keep punctuation. And we can also say why we need punctuation because it is necessary for the computer in order to be capable of distinguishing turns. And that means that we need to have a data like data representation that includes the configurations, we need multi grams, where punctuation is wrong. So the computer analysis, so once we have this kind of theory, that is backed up by experience, that we can make smart choices here. But as long as we don't, then the thing we're left with is playing around doing 15 variations that seem plausible ish to us. Choosing the one that looks good. And that's a sucky way of doing things.

32:22

Great. Yeah. Then coming to my last question. Also targeted towards the this will be second paper of my dissertation. So for example, in psychology, when you develop a scale that like specific, like standards on how to like, which tests to, to develop, which validation aspects to demonstrate that, for example, it gives us the scale will be put into the database here. And what do you think prevents the establishment of standards or frameworks within the research community? Both maybe about like, like more broader standards on maybe reporting people setting decisions also, and also about reporting specific dimensions of validity or specific validation evidence?

33:14

I think there's a few things. Let me say first, I'm not sure that this can be transferred in this form. Because sorry to say, but text is difficult, unlike surveys, by psychologists of cheating, because what they're doing is that basically, they approach somebody they asked him a question, and they called whatever what response they get. As if this was an answer to the question. That's an assumption. That's not to say thing, and that needs a basic you just need to deal with the thing that came back in response to the question or today isn't like this, like the text is not cooperating in your measurements. So you're asking the text, whether something is populous, but the text has absolutely zero intention telling you. And that means that for psychological measures, the nature of the construct, that you're trying to measure by scale doesn't really matter. Because you rely on the human brain of the person responding to kind of process organized and given a suitable response that is kind of pre structured in a way so you can put it in a scale for Tara methods really matters. I think it makes a huge difference, whether you're trying to measure whether the text mentions Obama, or whether you're trying to identify the three most common frames at any given time, over three centuries. And and that means depending depending on the kind of representation that the meaning that you wish to classify, can have an object different validation strategies. Maybe Rick Quiet. So I don't think that you can have a one size fits all set of standards, what you can probably do, and probably we should do is have standards for a number of well defined problems. And they, for instance, like one thing that comes to mind is that like after the F one, there must be an error analysis. That's, that's a standard, anything that should be agreeable. And it's a no brainer, obviously, that makes sense. But you can't really do an error analysis was an inductive pattern finding algorithms, right? Because there is no, no no ontological sense of arrow in this entire thing. So So that's, that's 111, big class of problems. And then the other part of the story is that and that's where the language skills that are now working on with objects come in, that there are a number of steps into data analysis that don't generalize. Because you only need them if you're working in Chinese or you only need to work we need them a few words in Hebrew. And the kind of strategies I mean, in the end, you need to get to some kind of harmonic some some standardized form of representing the text to a method, we'll work on it. But there is a set of pre process especially pre processing, but also some modeling problems, where the challenges that we face and analysis as substantively different in a way that it is not clear whether we can build a one size fit all standard for dealing with them. So like, last last thing here. Being the communication scientist, here, I liked the communication signs approach to this, which is to argue,

37:06

right? Because standards, standards are cool. If they are adequate, quick fix course, you probably aware of description of our

37:22

kind of debates about reliability. There are some people said, I guess, if everything that isn't criminals, alpha point seven just don't publish it is not the same. That assumes that there is a correct classification for every instance. That is not necessarily the case. So like, if you're trying to identify whether Obama is mentioned in a text, you couldn't have alpha should be one. Because we get that, but fuck you. But if you're trying to classify human and text, and that's something we do, it is not a correct assumption that for every textbook exists one correct classification. And therefore the assumption that underlies the standards that has been established in the field is problematic. And what it effectively achieves, is that there are certain kinds of problems that can only be studied in minor journals. Because if you study this kind of problem, you are not going to get this parameter IQ quality indicator that you need to publish in a major journal. And that's bad for science. So so I'm, I'm a big fan of standards where they are adequate. But I'm also a big fan of being able to argue your way out of standards and doing different validation strategies, because you have good and demonstrable reasons for why the standards do not apply. And I think if we took a lead, took a book out of this and another that political science is much more economical in the sense and much more centered oriented than communications. I like to think in communication that it forces you to justify things. Even if there is a standard, you still need to justify why this standard applies. And that's a good thing.

39:17

All right. Is there anything else you would like to add to this to this discussion? Before I close the recording?

39:26

I already said quite a lot of things. I mean, I can continue on this topic for ages. I've noticed right? Because I'm really interested and this is something that I've spent quite some time with in the last years. But I think that that's that's really I think the most important things that I think about the state of validation and maybe maybe one positive thing, we are getting better. There is there is some perceptible movement towards systematic validation, not only in the sense of it measures Something that looks plausible ish, but in the sense of actually checking whether the labels that people put on tools actually are justified, right? Like, for instance, you might be aware there's been like, a little wave of papers recently that trashed sentiment tools, which I love because I hate sentiment tools. Because I think there is no such thing as sentiment sentiment is not a valid category. There's like the validity problems at the category level, there is no such thing as sentiments makes no sense to try and measure it. But that's my problem. But they kind of taking these tools and re examining what they're doing and really trying to identify, what can they do? What can they not do? What can they do only under certain circumstances? Topic, the question has arrived, and that's a good thing.

40:56

All right, then I was