Mueller

Tue, 9/13 4:48PM • 25:49

**SUMMARY KEYWORDS**

validation, coding, method, supervised classification, approaches, measure, analysis, validate, measurement, human, improving, validity, measurement error, concept, people, paper, run, classify, report, reviewers

00:00

And it's okay, I can also switch to English now. So yeah, thanks a lot. I mean, we've already talked about the goal of this video. And I would like to start about a project picking what is validity or measurement validity in for textual analysis to you.

00:24

Okay, yeah, so I don't have a textbook definition in my head. But for me, roughly speaking, what I understand when we talk about measurement validity or invalidity, more broadly, in the context of text analysis, is basically to test whether a method or which of the selected methods kind of corresponds to human judgment. And whether it actually corresponds to human judgment in such a way that it can, or then when we can be confident enough to use this method, rather, you know, reading and coding our concept by hand. And if the text analysis method or computational method kind of approaches, coding of humans, at the unit of analysis, where and until can talk about that a bit more, so that it reproduces human coding, then I would say it's, it's valid. And one thing that I think it's quite important here is that for social scientists, it's not about not necessarily about, you know, running the most fancy method and having f1 score or accuracy that took them two percentage points higher. And at the end of the day, this cutting edge methods take requires, you know, a university server and one week of work, and computation. computation. One, let's say one week of computation compared to an easier method, that's slightly worse. But then once you aggregate this, you basically get the same results. And I think that's something where computer scientists and social scientists are heading might be different perspective. Whereas for computer scientists, the prediction on the let's say, sentence level, or paragraph level, and improving that and make this as good as possible is maybe one of the main goals. Whereas I think, for social scientists, more important to get this right on the level of analysis. So let's say we are interested in measuring what you said here, polarization for each legislator in a given week or in a given month, then there can be misclassification and that's completely normal. That would also be misclassification or invalid, or some of some of the some of our observations will not be classified correctly. But as long as this measurement error is not systematic, sorry, it's not. It's not systematic. So as long as the measurement errors, and systematic are more or less unsystematic, then our measurement can still be valid. So there's always some noise involved. And I think it's really important to focus on a unit of analysis and compare that to human judgment. So there's often an aggregation going on.

03:24

Yeah. Okay. Yeah. Thanks. And how would you describe, like, broadly speaking, the current state of validation of textual methods within the research field of Political Science, political communication?

03:40

I think it's improving in the last few years, and it is improving, because reviewers and readers are focusing more on validation. So if you don't validate your let's say, classification, or scaling, or only validated very briefly, I think you won't be able to get your paper published in a top journal anymore. Because there's more awareness for validation. And therefore, I think it has has improved but it's still not not the way it should be. Because also validation requires a lot of work. So if you if you write a paper and you think God it's that's interesting results, let's submit this but you don't validate. I and I think many of my colleagues wouldn't want to accept this paper because we need to see where it corresponds to, to human judgment. So text analysis is actually quite difficult. Maybe not kind of running the code, or writing the code is gets easier and easier because it's getting more user friendly. But validating this you need to get your hands dirty and you need to find ways to cope Are your results to other measures? Let's say you have expert estimates or you have self reported estimates or whatever it is all you have to do human coding. And I think it's improving because reviewers put more attention on it as more maybe also some job requirements. And there are also some books, textbooks and papers that highlight the need for validation.

05:27

Did you think that there are like differences between different groups of methods? So speaking about like this broad categorization of supervised unsupervised and blue based methods, that there are there's like, more advanced validation approaches or techniques for different groups of methods.

05:49

So I think that's the difference between supervised and unsupervised scaling on classification. So let's say, supervised classification. The advantage is that it's very honest to you, because you will see very quickly whether method works or not, so whether it corresponds to your head or test set. That can be good. And it can be also challenging, because you will always get an honest answer. And even if there are some miss classification, again, if on the unit of analysis that we are, let's say running our regression model, which might not be an individual sentence ban, which might be a manifesto, these hours cancel each other out. I think it's it's acceptable. If it's not, you know, if it's not as good as maybe it could be an all instead that also there's when we think about supervised classification, there's also a measurement error for our coders. For instance, if the same road as call centers, again, there wouldn't be measurement error. So ideally, we have more than one coder for our tests and training sets and compare inter coder reliability first, maybe check, okay, how does it correspond with we say, one powder cases to a given category, or we say, both corners, or we take three corners out, okay to certain texts or documents to the given category? What I find more problematic validations of topic models. So first of all, I think there's still a problem. And this probably won't go away. That topic models are great for exploration, but not very good when you actually know the concept you're interested in. So if you want to understand something about the economy or environment, I would always say try to use supervised method because you know, the concept of interest rather than using unsupervised methods, like atomic models, and then it changes depending on the number of topics depending on your pre processing, and so on and so forth. And I think often this topic model, validation is less transparent. And I'm not sure where they you know, you pick something that might fit your theory, but it's, you don't have really the performance metrics. So maybe you see a spike in one topic heading in time to go Okay, great. That's, that's exactly what I expected. So, it's valid or you select one or two documents and read them topic model validation requires a lot of work writing the topic models does not require that much work, but the Ex Post validations after you have run your topic model requires considerable time and I think here is a difference that researchers often let it not often, but are more likely than not to pay so much attention to the validations step in Topic models, because they have the results already in the thing Okay, that seems to work. Whereas for supervised classification, you need more thinking before you actually run your classifier. So, generally, I think the validation efforts are more on tend to be more thorough in supervised classification.

09:14

So, when you think about this, like classical research process from you have an idea then you read the manuscript, do some exploratory analysis, and then it's about writing the most probably a paper would you say that there are certain considerations or things to think about, about validation, which will not be reported or which is not talked about at the final paper level. So, if you have been published a paper and there are like some some some validation steps, but like some considerations or some tests maybe or some practices which you conduct with as a researcher, but which will not be like reported in the paper

10:01

It's hard to say for someone, I think that it's important and vital that all replication materials are published since, ideally, if the data are not sensitive, and I think for most political science data, okay, if we look at social media, maybe not. But for foreigners, it is possible to upload also your test training set so people could check on doing. And I think that's the best approach. So I know, for instance, for some of my papers, I run supervised classification, and it works quite well. But there's room for improvement. So I submit all my materials, and other people can use that training set, to run better classifier to classify us to extend it to investigate it. And by doing this, you're also contributing to the idea of providing good service to the research community can use and improve your data, and it also encourages you to spend as much time as possible on validation. But the first question was, for you to do all of this, do you have any considerations? And yeah, I would say it's, again, looking at humans, if humans are not able to distinguish between other to identify certain concepts, or a certain category. So that we need inter coder reliability tests and reading just what we are the reading text we're working with. And if humans are not able to identify a concept in most applications, not all of them, maybe if they are some latent traits, or money than human emotions, I know it's difficult, but if it's completely impossible to for humans to identify a concept of interest, then the our text analysis approaches, we're probably also spectral. Therefore, I think it's important to start reading the text first and trying to classify it manually. And then you might realize that it's not possible at all, or you might need broader categories or in B to follow different approach. And this is how to report in advance, you know, that you can report into coronary liability, and I think that's very good. You can have more than one coder. Again, if it's not supervised, as became an inevitable unsupervised classification, some reading of the concept of interest or some identification is, is warranted. So long story short if I don't know exactly how we can report this first step, but it's important before you run your your your models that you try yourself and other researchers or maybe call us try to identify the concept.

12:51

I think this does this does a good bridge to the to the last section of like this outlook thing, because, I mean, I also discussed this discussion with others. And some were pointing to this concept of semantic validity. So this to bring this human understanding of texts to those like other validation steps, which which are typically reported like this, convergent validity, you are criterion validity, content validity, maybe maybe that maybe that's connected to that. So yeah, maybe more broadly speaking, what do you think needs to be done? So? Are there any, like practices, which should be adjusted more like on the practical side? Or do we also have to think and like, that's the first question. And the second? Do we? Do we maybe need some, like theoretical new concepts of validity or some of our validation? Be more precise? Do you have any ideas on how to further improve validation of text data?

14:00

So good question. So do we need more concepts? I think the theoretical, as far as I understand the theoretical literature on validation is quite advanced already. It has been, you know, there's so much so much has been published and studied about validation approaches and different concepts. The problem is that maybe not all of that, maybe validation isn't sufficient. So I think from the theoretical side, we have quite a good idea already, what constitute our what, how to define and measure validation. It's more about on the empirical side, whether people actually do it. So you can come up with a new, new, more systematic overview of different validation steps, but I think there's quite a lot out there already. It's more about getting your hands dirty and actually validate Doing your mailers. And the best way of doing this going forward is first provide made some some some software that make it easy to validate or, you know, open source coding software that you can combine with R or Python. I think there are already some, some tools out there. But sometimes you need to have, let's say, our studio installed. But if you have trained coders who don't have our studio, how do you do that. So maybe some software limitations for straightforward validation approaches that don't dry out, don't require any technical knowledge actually set things up. That way, for validation purposes, I usually just use Google Sheets, because everyone can use it. It's online, you don't need to install anything. You have your documents, you have the categories, you will read through it. There's no technical issue. So that's that's the first point that you make it as easy as possible for yourself and other people to validate it. And the second point is that for, you know, that reviewers are skeptical of approaches that are not validated sufficiently. And then reviewers, or supervisors, or colleagues also make constructive suggestions how we could validate this. So instead of saying, I'm not convinced of their validation, say why, and how this could be improved. So at the end of the day, I think science is a collaborative effort. Of course, you publish yourself, but if you submit your application materials or production materials, code, data, code data, other people can use it. And on the other hand, if if people are reading your work, ideally, they make some constructive suggestions. So it's not about destroying other people's work, but thinking creatively and constructively. about ways of improving validation efforts.

17:07

Come back to those frameworks. Do you have any like examples for that? Because, like, generally, I get the feeling or like, I get the experience that there's like a lot of like theory about like, validate validity in a broader sense. But when it comes to the application, and I get the feeling that there isn't really this like, commonly accepted framework of standards. And this leads to the problem, in my opinion, that you can basically cherry pick the metrics or the validation steps, which maybe support your argument. And you can, you could like theoretically, leave out validation steps which are not supporting your, your, your argument?

17:52

Yeah, again, it's about honesty. If I were doing this, I probably wouldn't fear to. But maybe that differs across research. So I try to be always as honest as possible and transparent as possible. One way of doing this is providing, I think, various methods and comparing them. So let's say one method doesn't work perfectly. They're systematic SRE, there's this unsystematic measurement, measurement error or mis classification, then try a different method, and run your analysis with that method, which also has its problem. And then use the third method, what you are describing for your master's thesis, and sometimes this results, and it could lead to different findings. And maybe there are people who cherry pick one of the methods. I think this is a very just honest approach. I'm not sure how to solve this. But again, providing data and code could help other researchers to, to maybe apply a different method. And just some examples from my work, how I tried this. So let's say I want to identify broken unfulfilled promises. Using a dictionary approach works worked quite well. But also use the sentiment analysis. As an alternative measure of kind of maybe how newspapers report on promises. It's a different measure results were the same. You could also think about, okay, anything, if it's about victory analysis of supervised classification, tried to hand code. I tried to develop a directory, not by hand coding, but just by your domain knowledge, also do supervised classification and present results for both measures. So that's not necessarily about validation, and there might be validation steps that are left out, but if you can use different methods and come to similar results, they won't be the exact same thing To be seminar, that's that's one way of convincing yourself. And also, of course to the readers that, that you're that you're that you're having a valid measurement and that you can actually measure the concept of interest. And if it's, I'm sure if it's a very if there's an influential study other people, and you provide the data, people will, will probably look at it. Even if it's only one person, maybe this person has a better idea to do how to do it. And that's great. Maybe that person realizes that you could also do it differently. So again, Transparency is key, I think and that will, that will have that will have the researchers conducting the research to think about, okay, how is good and what can I do, and it also helps the community because these data are available, and you can make progress. So for instance, one last example in another paper, and we'll get our party story about the past, present, and future. And I use supervised classification. And because I provided all the training set and tested, I acknowledged that this is not working perfectly, but it doesn't seem to be systematic that measurement error, I don't do that by aggregating my results and comparing that to human coding. And I know of other people who are now using these data to develop a semi supervised approach. And they can compare that. And that's, that's nice. That's nice. And it's not that I'm that I don't want this. Like, I'm very happy that other people do this. And that, but that can only work if you're transparent. So I think transparency is key, because then also other people can check whether a measurement is about it.

21:49

Yeah, yeah, I also recognize that both of your papers are like, especially like this coalition, or like the scribing paper from the newspaper article, because I'm coding, like I studies relation to like some goal metrics, but also like, obviously, validation steps, and also if the data and code is available. I'm also quite interested in to see the final distribution of studies, who's doing that. Maybe one last question, if that's okay, for you, it's, it's 31. It's just Just a quick question related to the problems of how to deal with it, if you see that your measurement is not working, because there was one really important and great point brought up by another interview, that if you if you run, for example, and analysis, so your previous previous register your analysis, and then you see okay, there's no effect on like, no, in a relationship between your variables, that that's totally fine, because then it's still within your I mean, that's the knowledge you generate. That sense. But if you see that, you want to measure one construct, and you see, okay, there is no way on how to correctly measure it. Do you see any solution to this? Because obviously, you're not generating? I mean, you need to completely reframe your research question about maybe how to measure does construct them, but you won't be able to, to to assess your initial research question about like a causal effect between maybe one affected between one construct and another one.

23:32

Problems, I think if you're not able to measure it, then you need to be honest and not use that measure. First of all, which is which, which relates back to my previous point. I don't know how to do it. So one way, of course, is talking to your colleagues, talking to friends, talking at conferences, or workshops about what you're trying to measure. Maybe other people have have good ideas. Usually, community key is quite supportive. And this movement towards open science, open source packages, and so on. I think you're very welcome development. And you can talk about that, and maybe other people have have ideas. That's one way. I mean, you could also think about, I don't know, having some sort of a platform where you can, so how you want to measure something and why it didn't work. I don't think this will be too successful. So sometimes, some some concepts, I don't know might be too hard at the current stage for us to measure or to measure with quantitative text analysis, computational social science, but maybe it is possible to do it through human coding, I don't know. And that could always be way so I think computational social science and text analysis is great for scaling up your analysis. If you have a valid measurement But if the if your if your methods or your computational methods are not able to reproduce them with coding, maybe then switch back to human coding and select the smaller scale like a smaller sample and do this very thoroughly and do this in a good way rather than just trying to classify all data that's ever been produced.

25:26

Yeah, great. Thanks a lot. Is there is there anything else you would like to to add?

25:33

No, no, I think it's a very good research project PhD project. And I'm also interested in seeing the results and findings of your work and happy to help.

25:44

Yeah, great. I was stopped the recording right now.