Bot or Human? Unreliable Automatic Bot Detection

April 7. 2020

Hi, all, and welcome to this virtual Berkman Klein talk series. And I still see some participants joining, so I'll take it slowly here. But, first of all, welcome all to this series and this kind of like unusual setting, at least for the Berkman Klein Center in these very unusual times.

On your screen, you can see some of the house rules that we have. Most notably, audio and video have been turned off for you, so you will only be able to see the slides as well as the presenters, that is, Adrian and me. If you've got questions during the talk, submit them through our-- through Zoom's Q&A tool, and we'll basically go through those questions and answer them after the talk.

And after the talk, as well, the public chat will be enabled. So, right now, you can't be-- you know, you can't chat with other folks. But once the talk is over, you'll be able to do that more interactively, as we've kind of like learned in the past that this can be a little distracting if the public chat will be enabled throughout.

And, finally, a word of caution. As usual, with these talks for Berkman, the webinar will be recorded. And so, you know, like, if your questions are being asked, and we might mention your name, so keep that in mind. Or, you know, if you don't want your name mentioned, just add that to your question, please.

With that having said, I'll now share my screen with you. And hello, all, in this virtual environment.

Today, Adrian and I will talk about our research on bot detection or, more specifically, on the issues of bot detection. And so this talk, as follows, is structured in a way that, first, I will guide you through kind of like our thoughts and how we kind of like structured our paper, what kind of like the situation is that we're in right now when it comes to bot detection, and then Adrian will guide you through our results and give you kind of like some thoughts on what we think can be done and maybe should be done.

But, first and foremost, we are just very happy that you chose to join us online. And, you know, like these times are quite stressful and uncertain, and we're very thankful that you're participating here. And I hope you all are safe out there.

So when it comes to, you know-- and this is a very abrupt change-- but when it comes to bot detection, which is a very more like abstract topic, we kind of like have to take a step back and think about the public interest, or worry, really, in bots and how this has increased over time. You can see the plot on the right. That is from Media Cloud. So this is just media attention in the

US media between January 2015 and April 2020 for bots and fake news, misinformation, or disinformation.

And this kind of like just shows you that, really, after fall 2016, so after the presidential election in the US, the public attention and probably worry about bots and kind of like misinformation really, really skyrocketed. And this is not only true for kind of like the public and the media interest, but really also for academic research.

So Google Scholar has 18,900 hits for me when I search for bots and Twitter since 2016. I couldn't use Scopus because, as many of you are, I'm working remotely, and so I had to take Google Scholar as my database of reference. And so this just shows you kind of like that this field of bot research and kind of like what-- what's the role of bots in our public discourse, it's an urgent one. Right?

I think no one is disputing that we all would love to know kind of like how many automated accounts that spread political messages, or other messages, are really out there. You know, like we all want to know if we debate with someone online, whether that person is a person or really a bot. Right? So there is a need and an urgency for that.

And, indeed, you know, like there have been studies in recent years and articles about this. So this was, for example, a Pew study from 2018 called "Bots in the Twittersphere," and where they say an estimated 2/3 of tweeted links to popular websites are posted by automated accounts, not human beings. Another study that was covered by The Guardian, this year, had the headline, "Revealed Quarter of All Tweets About Climate Crisis Produced by Bots."

This, I think it was by Gizmodo. This is an op ed which says, "Social media bots are damaging our democracy. On the internet, nobody knows you're a natural language processing system." So, again, or really, you know, like it's not a worry-- it's not only a worry, it's also-- it might be really bad for our democracy. Right?

And finally, you know, like all different kinds of things, "Tweets About Cannabis' Health Benefits Are Full of Mistruths." This was I believe on The Conversation about a Twitter analysis that some academics have done.

And this is kind of like the interface that we're kind of like looking at, between academics doing research and kind of like looking at discourses on Twitter and being like, OK, how can we measure this? What's happening there? Then writing a study, and then kind of like this, getting out into the media and the public discussing about this. Right?

And so the importance of the rigor of the academic research is, you know, warranted in that regard because, often, kind of like academic discussions and studies are usually staying within academia here. It's not only in academia. It's also out in the world and out in the media.

So when we talk about bots, we kind of like have to think about the terminology, and we just, like, have some definitions here, for example, by Chu et al who say bots are automated programs, but they kind of like also differentiate between bots and cyborgs, that is, bot-assisted

humans or human-assisted bots. That is kind of like, think of a Twitter program, and sometimes a human will also write tweets on that account.

But this is only one of many definitions that you'll find out there when you kind of like look in the literature. Another is by Kollanyi, Howard, and Woolley, from the OAI, back then, at least, who defined bots as, "automated interaction with other users." So this is already kind like it's about the interaction with other users. It's not only like, you know, like that they send messages into the vortex.

Third, it's from Bessi and Ferrara, who not talk about bots, but about social bots, who, "emulate the activity of human users but operate at a much higher pace, while successfully keeping the artificial identity undisclosed." This, again, is a different definition, and you kind of like can see that from definition to definition, the line and the lines of what is a bot and what's not a bot beginning to blur.

And, finally, Bot Sentinel is a website out there that came up with the term trollbot, which they define as, "troll-like behavior with a repetitive bot-like nature of their trolling," where, you know, like it's really unclear if that's even a bot or not. And, you know, like, in that kind of sense, does it even matter if it's a bot or not?

And so it, kind of like you see that definition, right, so you kind of like have a broad feel. And so we believe that it's kind of like important to, you know, take a step back and say, for us, bots are fully automated accounts. And that's, for us, it's the end of it, not because we do not account and believe that there are different ways, out there, that bots are being used, but rather because we don't believe that, like this, all these terminologies help us kind of like figuring out what we want to study. That is, you know, how many bots can we identify on Twitter?

And we, when I talk about bots, you know, like there are some very obvious bots out there like the Museum Bot or the Soviet Art Bot often accounts, you know, like that have their engineer in the description and a link to the GitHub repo. So how do you identify bots? Or how have scholars and journalists and researchers thought about going about this?

There are different ways. There is the frequentist approach, which is really just saying, you know, every account that tweets more than 50 times a day is a bot. There is a network analysis approach, where we kind of like look at the follower communities.

There is a machine learning algorithm, where you say, OK, we know these accounts are bots, and these accounts are humans. And we train this algorithm so that it will detect these, going forward. There's digital forensics, which is kind of like a mixture of human inspection and these other tools. And, finally, human inspection, where it's really just, you look at the account and you look at what they tweet, and you kind of like try to make sense. You know, like does the image fit, et cetera?

So here, in our talk, we focus on machine learning algorithms, most namely Botometer, which can be understood as the gold standard in social science research to identify bots on Twitter. Indeed, if you kind of like look through the literature, you'll find several hundred studies that

have used Botometer, including and kind of like communication journals like Political Communication, but also, for example, in general science journals like Science. And so, you know, Botometer is, by far, the most used tool in that regard. And so we believe it's valid kind of like to ask, you know, how good is the Botometer?

And I went through Botometer with these two bots that I just showed you, and the Botometer score, here, is 2.4 and 2.3. So, you know, it's in the middle. Like, Botometer is kind of like saying, you know, it depends. The whole left, like blue, is not a bot. Red is bot. So it's right in the middle.

If you're wondering, and if you've never used Botometer, how it works, this is the web interface. And you log in with your Twitter credentials, and then you can just check users. And what you'll get to see, here, is what I've done yesterday. And I checked Adrian's and my account. And apparently, Adrian is a little [? bottier ?] than me. But, still, I think we're both relatively in the clear and tend not to be a bot, according to Botometer.

So what you see, here, are the universal scores. You can see this also on the next slide, where I looked at Adrian's account and where we can get two important values. That is the Complete Automation Probability, which is 1%, and the universal score. These are the two main values that you will also get from the Botometer API, and with-- and those are the values with which researchers usually work.

Most prominently, however, researchers will take the universal score which is then resampled to zero to one, where zero is not a bot, one is bot, and they will choose thresholds. And above or below those thresholds, it will either be a bot or not a bot. So Pew, for example, had a threshold of around like, I think, 0.43. And everything below 0.43 was not a bot, and everything above 0.43 was a bot. And that is the universal score in that context.

And if you talk about these classifiers-- and Botometer is a classifier, right-- we talk about pitfalls and what can go wrong. And we see this currently, actually, with the coronavirus and testing, where it's really important how good is the test. Right? And this is very similar with classifiers. Like how good is this, actually?

And so if we have a sample of 50% bots and 50% human users, the question is what pitfalls can there happen? Generally speaking, you know, you will have a classified data set. And within that classified data set, you will have true positives-- that is, in our case, the bots that we identified are bots-- and you will have false positives. That is, you will have humans that have been classified as bots, but obviously aren't. And usually, you know, like, this is something that you will never really like get rid of. But, obviously, kind of like we want to reduce the false positives and have a high true-positive rate.

However, an issue arises, here. That is, that is not the only pitfall that you have because true positives and false positives only tell parts of the story. Indeed, it's also a precision. That is, from the poll, the accounts that got classified, the question is, how many of these classifieds were actually correctly classified? Like, how many bots were selected within the classified data. That is precision.

And we have got the recall. So how many bots were actually selected, correctly selected, from all the bots that we have in our data set? So when we think, and going forward for this talk, it's important to differentiate between true positives, false positives, precision, and recall. And you'll find out later why that is and why we can like talk about this.

But, more specifically, because there is an issue if we only look at true positives and false positives. Namely, that this data set doesn't really happen on Twitter because it rather looks like this where we have about 15% of all users, according to Twitter, are bots, and the rest are humans. And so this is not really being accounted for, and you'll see the impact of that.

So this is all theory, and this is kind of like all the aspects that we thought about in the last one and a half years while doing this project, just coming up, really, with this question. Like, how precise is this tool in detecting bots? And we've got four bigger questions and then a smaller question in that regard, I think.

That is, how good is the diagnostic ability of Botometer when used for five distinct sets of Twitter accounts? How good is the precision at the recall of Botometer scores when used for five distinct sets of Twitter accounts presenting the bot-human ratio and the general Twitter population. So not only in our data sets, but also when we resample this for the general Twitter population, which is kind of like what we're really interested in, right.

But, also, what's the difference between the languages here? We know roughly that there is a difference between how good, for example, Botometer works in English and Swedish, based on other studies.

And finally, how stable is Botometer over time? Because usually how studies go about it is they collect their data, they run their data through the Botometer API, and one time, and then they have the results. And that's usually it. And we're kind of like also interested in, how do these values change over time?

So what we did is we constructed a data set of five different data sets. Namely, we had clear humans, clear bots, and a training set from Botometer. The humans were US politicians. So here's some examples, here, usually verified. We have German politicians, mostly verified. But all, you know, like all accounts that were clearly humans and the media would notify the public if something really weird and fishy would go on, that is, automation in one or the other regard.

And, finally, we also looked at bots, right. We went to a wiki that has Twitter bots on them. Those are usually, you know, very transparent Twitter bots. So if we were to do a human inspection, we would all be like, oh, yeah, that's a bot. Now, and we did this for new bot, for new English language bots as well as German bots. And then, as I said, the bots that were used to train Botometer initially.

We had, then, this data set of 400-- or 4,000-something accounts that we then queried Botometer with. And we queried Botometer for three months. So we kind of like wanted to really have like a lot of reference scores to then calculate our scores. And, with that, I'll now let Adrian guide you through the results.

All right. Many thanks, Jonas. So let me start with the five different data sets that we have created for our analysis. Because Jonas has already introduced the single data sets, and to really test a classifier, a binary classifier, we have to, of course, use data sets that have both bots and human accounts.

So the first data set we used is like just all the accounts we have identified together. That's here, red, the all. Then we combined the German politicians and the German bots, but we only had identified very few German bots. That's why we created, also, like a third combined data set with the German politicians together with the English bots. And then we have a fourth data set, the US politicians together with the English-language bots. And then, as a fifth data set, we used the data set created by the first-- that was used for the-- to kind of like train the classifier by the creators of Botometer.

So usually when you want to report or analyze the diagnostic ability of a classifier, you report something called the ROC curve, which stands for Receiver Operating Characteristics, and then you calculate the area under the curve. So this means like the area that is covered by the curve, what follows under it. So the larger that area, the better, actually, the classifier.

So let me start. What you usually plot, it's like for every single threshold. And, for us, the Botometer score goes from zero to one. So we start on the right-hand side with a Botometer threshold of zero. So if you say your threshold is zero, and every account that gets a score above zero, you classify as a bot, you get a perfect true positive rate. This means that you will identify all bots within your data set.

But at the same time, of course, every single human user in your data set will also be classified wrongly as a bot, which means you get a perfect false positive rate of one. That's why you start on the right-hand side, in the upper corner, with these visualizations.

Then, on the other hand, if you move up with the threshold, you increase the threshold, then eventually, of course, you can use the highest possible threshold of just below one, maybe. And then what you get is actually a true positive rate that maybe includes only one bot, so below even 1%. But at the same time, you also reduce, of course, the false positive rate because no human account will be classified as a bot, wrongly classified as a bot. But at the same time, you won't really identify, also, any of the bots in your data.

So we compare, here, in this visualization, the different data sets. And we used, actually, the mean over the three months. So, everyday, we measured once the score for every account. And for this part of the analysis, we just took the mean, the average score, that an account received over the three months.

And what you can see here already that the US politicians and bots curve, the ROC curve for that data set, is better than the other curves. As you can see, the German politicians and bots as well as the German politicians and German bots, they are a little bit worse. The area under the curve is actually smaller, and you see that here as a summary.

This is usually what is also recorded in studies and also the Botometer creator report, the ROC in their paper, and it's pretty high, actually, in their original paper. And, here, you see the US politicians and bots, they really receive the highest ROC area under the curve score, whereas the German bots and the German politicians and the German politicians with the English-language bots together get a rather low overall ROC-AUC score of 0.76 and 0.77.

So the problem is, like with the ROC-AUC approach, you get like relative values. You get a percentage, basically, for a single part of the data set. Like, people get once measurement for the bots, and you get another one only based on the human accounts.

So Jonas has already mentioned it. When we use this classifier, our target population is not our training data set or an artificial data set that is balanced. In reality, actually, we want to analyze the Twitter population. And within the Twitter population, we can assume-- I think we all can agree on that-- that we have less bots than human users. And, of course, there is no clear number for that, but we find numbers like 15 percent of the accounts active on Twitter are actually bots.

So we try to create a new data set that has this kind of imbalance, because that imbalanced data set is what you will actually see in the real world when we analyze Twitter outside of your experimental setting or test setting. So we created a new random sample with replacement, with 100,000. Each data set, we created a new version of it with like 100,000 accounts, but we adjusted the probability weights during the sampling process that, in the end, we get basically a data set with 15,000 bots and 85,000 human accounts.

But I told you already, before, the ROC approach actually gives you relative values. So no matter how imbalanced the data set is, as long as you use more or less the same population or the same kind of accounts, you will actually get the same ROC-AUC score. Here, on the left-hand side, we see the original one based on the training or our data sets. Most of them are actually balanced. And then, on the right-hand side, here, which is like labeled as sample, you see the new data sets we created with the imbalance, with only 15% as bots. So the scores, here, are overall the same.

However, if you use as a measurement, actually, the precision which analyzes-- like, it shows here with the imbalanced one-- it analyzes like from all the identified accounts that are above the threshold. How many of them are actually bots, in our case, right? And if it's imbalanced data set, even a very small false-positive rate or overall false-positive rate will lead to a high absolute number of wrongly classified human accounts that are suddenly classified as bots. And the number of bots is actually smaller within this identified part.

So what does it mean for the PR, the Precision Recall, area under the curve? On the left-hand side, you see the PR scores for the original data sets which, of course, in most cases, is better because the data actually is more balanced. And then, in contrast to that, on the right-hand side, you see the PR score for our newly created imbalanced data sets.

The only value that actually increases, what you can see here, is the German politicians and German bots. The PR score, here, is smaller for the original one because we had very few German bots. So we had less than 15% overall in that data set, so when we created a new data

set, actually the score increased because we had to add more bots. But for all the other values, actually it gets worse.

So let me show you, visually, what that means. So if we compare the two PR curves, the one for the original data sets and the other one for the resample, for the even newly created imbalanced data set, then you can see, here, for example, for the data set All, the curve gets pushed down. So the precision, that's really the value we're interested in, gets worse action.

The curves look more or less the same because we sampled from the same population. And the recall, of course, stays the same. But you see, here, for the German politicians and the bots on the right-hand side that it really pushed down this curve.

I will, now, explain what it really means and how you can interpret these curves on the next slide, here. So these curves read a little bit differently than the ROC curves. You started, really, on the right-hand side with the Botometer threshold of zero and, on the left-hand side, you plot, actually, the thresholds that are the highest. So from zero, on the right-hand side, up to the left-hand side, with a threshold of one.

So, again, if we take the lowest possible threshold, and we say like, every account that gets the bottom score of at least above zero, right, we will classify that account as a bot. We actually get the proportion of the population. And as I have told you before, now, in our newly created data set, we have 15% bots and 85% human users. So the precision is exactly 15% because we get all the bots. But, at the same time, we also get all the humans, so we have 85% actually wrongly classified accounts.

So we plotted, also, into this visualization, points that show you thresholds that were used in already-published studies. So the ones more in the middle are actually the points. These points show you the threshold of 0.43. That's the threshold that Jonas has already mentioned before for the Pew study.

And if you take, for example, the data set All, you can see, actually, the precision is not that great. It's actually 0.5. So if you say, all accounts that have a score higher than 0.43, the identified accounts, 50% of them actually will be bots. But the other 50% will be false positive, stable, human users.

If you do the same for the German politicians and bots data set, as you can see, it performs a lot worse. It's even less, actually. It's more in the 30% range of like the number of spots that you will identify, and more like over 60% are German users.

The points on the left-hand side are the thresholds for 0.76. These were used in the German context in a study that is published in Political Communication. And, as you can see, there, it's a rather conservative threshold. Right, there's pretty high one, 0.76.

You still, you still get like a lot of, like, wrongly classified human users into the data, and only around 25% of the as-bot classified acounts are really bots. And at the same time-- and this is the recall, right, that you can see here on the x-axis-- the recall is very low if we use a very

conservative threshold. So conservative, here, means a very high one where you, of course, want to reduce the false positives.

But when you use a very high threshold, at the same time, you will identify very few bots from the overall population of bots within your data set. So what this visualization also shows you, there is, obviously, a visual difference between the quality of classification between the English language, more general corpus, versus the German corpus, which is, here, a lot worse.

So here is a summary, all the scores, and what is, here, interesting for us, actually, is that last row with the PR scores. And you can see the German data sets perform a lot worse with Botometer, whereas the English-language data sets all have a higher score for the ROC as well as the PR. These differences are also significant. You can read more about this in the paper. So there's also a test to really check whether it's significant.

So as a last step, we also wanted to find out whether Botometer gives stable scores, whether there is like a low volatility or a high volatility. And we have like selected a few of the accounts. So, here, we have a far-right politician from the Alternative for Germany, Alice Weidel. And you see, here, over the three months, the scores plotted that we measured on every day, and there's a lot of volatility.

And if you take the threshold from the Pew study, where you say, like, every account that gets a score higher than 0.43 is a bot, actually, you're just at the end of March. She would have-- she wouldn't have been in a bot. Whereas, on other days, actually, during these three months, she would have been classified as a human, but there is a lot of volatility.

So let's go to the next one. This is like an interesting one. And a US politician, actually, a Republican, has a very low bot Botometer score in the beginning, but then it increases. And towards the end, also, again based on the Pew study thresholds, he would have been classified as a bot.

And as a last one, we have, of course, chosen a bot. This is-- Jonas has (LAUGHING) identified this one. It's the Boston Snowbot, and as you can see, here, obviously, the Boston Snowbot is tweeting when there is snow. And it seems like, in March-- I don't know Boston well-- there was some snow. But then towards the-- towards April, and then starting from April, somehow, probably this bot wasn't really active anymore. We think, here, also like Botometer captured probably the activity level, and he was far more active in March than in April, and maybe he even stopped in April.

So what we see here, with these very few cases already, there is a lot of volatility. And if you think about most bot studies, they only measure the bot score once on a specific day, and not like over time and try to take an average. And this really can be an issue depending on the day, on the count will be a bot with a certain threshold. Then, on another day, he won't be a bot.

But we also have like a summary. This one, you'll find, also, in the paper where we really check over this, over the three months. And for every single threshold, actually, whether an account was at least once above and once below the threshold with the score.

And if you take, here, for example on the left-hand side with the threshold universal score, a threshold of 0.4, which is very near, again, to the Pew study, then you'll see over 55%-- this is the red line-- of the new bots, of the English-language bots we identified, they had at least once, on a day, a score above that threshold of 0.4 and, once, during the three months, at least once, a score below the threshold. So as you can see, the same holds true for the German bots.

With the human users, it really depends. It's maybe less problematic, but still pretty high. For example, you can see, here, the German politicians, again, on the left-hand side. If we take a threshold of 0.4, we still get around 30% of the users that at least have, once, a score above the threshold and, once, a score below the threshold during these three months.

All right, then we're at the end. And so what's the conclusion? What are the learnings, here? We should, of course, be very careful when we're using Botometer because the scores are not stable. There are problems with the language. Maybe it works with English. It doesn't mean it works in other languages.

And then what we believe is really our kind of like strongest message, it's like we really should consider the even balanced sample, right? So the populations we're analyzing with a classifier, the population maybe is not as balanced as our training data. And this is a general problem. Also, if you read the literature about classifiers in bioinformatics, there are a lot of discussions about that.

And if tests are developed and then used in a real population, so you also have-- and this one we haven't mentioned before-- you have to consider what is more important or what is more costly, false positives or false negatives? But this is also something beyond our paper, but definitely a question that every single researcher, when he approaches the problem of bots, should answer.

So how should we go forward? We recommend, of course, as communication scientists, always do some manual validation. Just because a classifier was validated in a prior study doesn't mean it still works with the new data set or new population.

So, for example, in the Swedish case, some colleagues have tested Botometer. They got very bad scores, so the decision was that they created their own classifier for their specific Swedish data set, and then it worked. It got better ROC scores, but also better precision.

So you need, also, to validate over time. This is something very general. We think, also, in general, any kind of classifier that analyzes Twitter data or Twitter accounts, even maybe beyond just bots analysis, they should consider that if it's not like stable over time, then, of course, the validation in different languages faces more from a, let's say, European or Asian perspective, that it's, for us, may be obvious, but for a lot of US researchers, it may be less obvious when they create their classifier.

So, of course, if we, now, move forward and say Botometer didn't work well, we can just create a new classifier. It's very easy, now, based on this data set to almost create a perfect classifier. I mean, I can just create an ad hoc classifier where I check whether a bot is mentioned in the description or in the username. And then the classifier already works pretty well, actually.

But that is definitely not enough, even when there will be a new classifier available that maybe was validated when other researchers would use it. In the future, they have to validate it again, like we say. So if you're interested in that, please read our paper. We explain that in more detail, how to move ahead, but also all the technical parts are explained in more detail.

And, of course, what we also call for, we say these kind of blackbox tools where the code is not available and the full data set with all the measurements that was used to create these classifiers, if that one is not available, that's a problem. Right, we cannot really reproduce, one-to-one, exactly the same classifier. And we need that to evaluate, actually, the quality of the classifier, to really understand where the classifier is biased.

We can now only, somehow, reverse engineer it or do some guesswork based on single examples where there is a bias. So we also share, of course, maybe call for more transparency. We also share our data. You find the code in the data, in that Harvard data words. So that's it from my side. Many thanks for joining.

Thank you so much, Adrian. I'll now have the double function of being both basically host and asking questions as well, maybe answering some questions. And also, a special thank you to Adrian because he's in Taipei at National Taiwan University, and it's, oh, 12:40 AM there, right now. So thank you for staying up so late.

The first question, actually, reached us via email before the event had even started, so we want to give space for that, too. And the question is, what do you think are the prospects for passing federal regulation on bots in the near future, and how do you view state laws on bots being carried out? And as Adrian is an expert in that regard, I would want him to go first.

Expert? It is a very interesting question because, like, two years ago, before I came to Taipei, when I was still back in Switzerland, a group of Swiss Members of Parliament invited me to Bern, and we had a kind of like a closed-door discussion with some legal experts that have written, also, papers about like bot regulation or potential bot regulation. And they had this idea of, like, regulating bots because they read all these stories in the media, right, about like bots take over a democracy. Or it's like the biggest threat to democracy. And then we had this discussion and, even back then, I was rather skeptical.

So let me start, now, based on our analysis, now. We can say it's very difficult, first of all, from a conceptual perspective. Right? And I'm not a legal scholar, but you do need to be very clear what you mean by a bot, if you want to regulate that one, right, object. It's very difficult. Then, as a second point, it's very difficult to kind of like identify bots, like to measure it. Often, it's not really clear who is a bot, and there are like accounts that have some, a certain degree of automation, but then they also have humans behind it. So it's very difficult to identify them and also very difficult, at least at this point, to define it.

However, and this was really the last point that we were discussing there and, actually, the legal experts back then, at least the Swiss ones, they also agreed with me and other people involved in this discussion that probably we have to ask the question of, like, how big of a threat are bots in comparison to other threats to democracy? But also, like, what's the effect of these bots?

And what we can say, or a lot of political communication scholars say, more traditional political communication scholars, they say it's very difficult to change the opinion of people, in general, even like with strong campaign techniques. So, of course, if at all there is an effect, it will be a very small, probably a very small effect. And at the same time, if we think about foreign interference-- and, now, I'm speaking more from a Swiss perspective, but also I can speak from a Taiwanese perspective-- there was also this debate before the presidential election here in January. Like, the China threat, right, especially online warfare.

However, if you have a broader perspective, I would say the biggest threats, actually, are offline. So in the Taiwanese case, for example, there's the discussion of direct proxies. Probably that's far more effective. So this means you go over organizations or persons or even politicians. Right? Of course, that is happening, and if you want to fight foreign interference, you should start with the biggest issue or where there is the biggest threat. And I strongly believe bots are probably not the biggest threat.

At the same time. of course-- and this was also the conclusion in Bern, back then-- we should keep an eye on it. Maybe the situation, of course, is changing. There will be new developments, and maybe in the future, in a certain context, bots might be a problem. But, at the moment, it doesn't-- it's-- we were advising against regulation. And I think the same, I would still say, today. It's very difficult and I wonder how it's possible to regulate it. And then, at the same time, is it really the biggest issue we will cope with at the moment?

Thank you. And I think I don't really have much to add to that. My intuition is before and after doing this study, that if you do try to implement a regulation, that the limits of the regulation, in the end, will be so tight that you won't capture a lot of aspects, which then, in turn, just like asks the question, how effective will be the regulation, in the first place?

Plus, I don't-- you know, like I don't, in the end, see bots as kind of like the biggest problem, even when we think about misinformation.

A second question, then-- and this is maybe one for you, Adrian, as well-- is from Bao Bao, and she asked, "Could a classifier like Botometer use Bayesian methods to take into account the imbalance between bots and real humans on Twitter? They can use what's the latest breakdown of bots versus humans as their prior."

Theoretically, yes, you can-- you can even-- if it's more about like testing your classifier. So during the validation phase, you need to take that into consideration. So if someone tells you, here, I have a new classifier. It works perfectly well. It has a very high oral C-score, over 0.9.

And then, you want to use it-- of course, you should validate it. So you could validate that new classifier with your own data from the population. And when you use a data set to, again, validate it, then you need to be careful to not use a balanced data set, where you have like 50% bots and 50% human users. You can use that to train the classifier, to add more information, put more emphasis on one group.

But, to validate it, actually, in that one, you have seen, now, in the presentation. With the precision, you should take this imbalance into account. And then there, you have, of course, different methods how to test it, right? The Bayesian approach, itself, I'm a big fan of Bayesian regression models or, in general, of this paradigm or way of thinking.

But, ad hoc, I can't think of like how to use it. Of course you should always-- this is your idea right? Of course, the mindset is similar. We take into account. We have some priors, like we assume the population is not balanced. I would just say the validation gets better if you use such an approach. But if the classifier, then, itself is better, I don't know.

Awesome, thank you. The next question-- and this is, we have several methods questions and more general ones, and I'm just going to switch between those. And so the next one is from Rod, and he asks, or he first says, "Thanks for that fantastic perspective. Given that Oxford Internet Institutes research points to the increasing disinformation architecture of countries like India, Russia, and Iran. Can you point to work, if any, being done on multilingual sentiment analysis?"

Could I answer it?

Please, go ahead.

Very, again, it's exactly the same. I would say, if you're talking about sentiment analysis-- and let's even roll all the conceptual discussion about what is sentiment. Let's say you can really measure sentiment, right, and it works pretty well in one language.

If you want it with multi-language, and you were interested in the comparative perspective, where something is stronger or weaker, of course, you'll need to validate it, not only in every single language. You also have to check whether it's similar in every different language, the scale, right, and the strength of the sentiment.

Then I could start with the cultural question, whether these kind of like sentiments are really the same within a culture. Maybe a strong sentiment in Taiwanese culture, or expressed sentiment, is stronger than just a strong sentiment in the US culture, right, where it has to be even stronger to be considered extremely strong, right? But whereas, here in Taiwan, a strong sentiment would be already considered as an extremely strong one.

Again, validate. I also test, as always, with my students. Don't just take these out-of-the-box solutions without any validation. Some of my colleagues in communication science have tried to revalidate a lot of these word lists that are used for sentiment analysis. And, actually, if you don't do anything, you don't change these-- word lists and you try to validate them in a new context, they don't work at all.

So, as a baseline, usually-- and this comes more from publications science again-- you need to set a gold standard. The gold standard is usually the human coding, and then you're matching whether the human coding is the same as what's the sentiments classify or the sentiment analysis measures, whether there is like a strong overlap.

So between different languages, it's very difficult, but what I can tell you, if you have a list for a specific language, if you're just within one language, you need to change the word list for the specific context. It can work quite well, actually. But, again, you need to validate it. If it's not validated, I would be very skeptical.

Thank you. The next question is, really, I think like, what was it , like six minutes into the presentation? It's like, unreal. "We want to know if we are debating a person or a bot issue. Is it your sense that there are bots that can pass that particular version of the Turing test?"

And I'm just going to take that question myself, first. And I'm also excited to hear Adrian's opinion. From everything I've seen on Twitter, I would say no. I think a lot of these, you know, like Twitter just debates, especially are usually short. And so you don't have a lot of time, and I would say, also, like in more heated discussions, probably a lot of projection is going on. And I haven't seen anything that would kind like point me to a different opinion on that. What do you think, Adrian?

Yes, I agree with you. I can maybe add one more point from two German colleagues, Pascal Jurgens and Simon Kruschinski. They also analyzed astroturfing and automation on-- actually, on Facebook. And what they say, it is actually the other way around. You will be surprised at how many human users use very simple communication patterns that could be interpreted, actually, as automation or as like even like the text, it's like automatically created. They just use emoji.

Because the majority of people, actually, on Facebook because they had a very large data set of comments, and what they say, the majority of people they are not communicating like academics on Twitter when the have a debate. Most users, they write very short sentences, not even sentences, a combination of emoji, or even like wrongly spelled words, single words.

And they are still human users. So we need to be very careful, even the other way around, to kind of like be careful to not label human users as kind of like bots. And, in reality, they're actually a human users.

And with regards to the Turing test and the bots, as far as I know, there is not really a bot who has like passed the test. But maybe we will see, in the future, there will be a change

Thank you. The next question is more of the methods side. And the question is, what is the total sample size for each of the groups that is the n? I had that slide up in the middle of presentation. Overall, there were 4,400-something accounts. And I think the majority was 2,000-something accounts were from the Botometer training set, and the rest was, I think, roughly 50/50. Or, like, I think they were a little over 1,000 humans and, again, like a little over 1,000 bots.

And the second question, in that regard, is the size of--

Can I say one thing?

Yes.

Yeah, that's also a limitation we have to add.

Yes.

We have a very narrow data set, and we specifically have chosen this one. But at the same time, these very homogeneous groups are also a problem, kind of like any mutation. So maybe with other data sets, Botometer will work differently.

Yes, I think that's important to note, and also why you kind of like choose this data set was because we really wanted to be sure what we have in our data set, right, because otherwise it gets very like tricky. And so the second part of the question is, is the size of n more important than the level of imbalance of the data set? What is the impact of different values of n?

I mean, a too small n, of course, is this problematic. And then, again, I would say like it's more about like, is the general population homogeneous or heterogeneous? In which context do you want to use a classifier? Again with typical traditional sampling theory, but you can say the more heterogeneous your population that you want to analyze, the larger the n, also for your training data sets, to really gets a good classifier if you have a very, very homogeneous data set. Of course, you can even use a very small n And you will get a very good classifier

Thank you. The next question is from Maria, and she asks, "How does this compare with the R package tweet, bot or not.

Tweet, bot or not, which one is this one?

And I have to preface that answer with, that we haven't checked specifically other packages against this. We know that there are-- you know, like other options of bot detection out there, which all have their advantages and disadvantages. We picked Botometer specifically, since it's kind of like been used for most, like in most studies.

And Adrian has pulled up, I believe, the R package, right now.

Yeah, exactly. No, I know now, which it is. It really depends what Mike Kearney was using to train his classifier. And as far as I understand it, he used probably even the same length, but he can maybe tell more about this if you ask him. I think he has chosen, also, like these obvious spots. So, in that sense, I strongly believe that package will probably have a better performance than Botometer, if you use these accounts.

But you need to check, right? If you use the accounts that were used to train the classifier, of course, you will get an excellent performance. So, again, our recommendation is that even if you get an almost perfect classification with an artificial training data set, use it with the population you want to analyze.

And then validate, really, once you get the ear as like as spot-classified accounts. And then we check how many of them are actually false positives. And then, at the same time, and this is more difficult because usually the majority of accounts will not be classified as false, right? You

still need to check, and this is a lot more difficult because then you have a very large data set, right?

How many of the bots were wrongly classified, or not identified and wrongly classified as human accountants. So you need this kind of manual classification anyway.

Awesome, thank you. So we have seven more minutes, so we'll try to keep our answers short to get through all the questions. Because I was told we have until 1:05.

So Mason asks, "Are you aware of any use of bot classifiers by platforms as part of their content management strategies? And, if so, do you think these classifiers are doing more harm than good?" And I personally, you know, I haven't talked to anyone within the social media platforms about that. I would assume that they have classifiers that they use amongst others for content management.

You know, like obviously platforms don't want bots just like rampaging around on their platforms. And if, you know, I think if they don't invest also heavily in human eyeballs that check and validate the accounts that they're removing automatically, I think there might be some harm being done along the way. What do you think, Adrian?

Yeah, I totally agree with you. I mean, obviously, they are-- they have probably their own classifier, and they have like a lot more information available. Right? The only-- also, the creator of those platforms, they only have the information available that you actually see when you open a tweet or an account, basically what the API returns.

Twitter, of course, they have all this backend data when a user is logged in. The IP addresses, right, at this information we actually lack, and I think if you have that information, it's easier to follow. It's easier to create a good classifier.

But, again, probably what the platforms think about, they really think about this false positive rate. Right? So, again, there is really a cost if the false positive rate is too high, and you really say every account that is above a certain threshold, you kind of block them, right? And you block too many real users, of course, it will have like consequences, and people become aware of it.

So I think they are, rather conservative with this. What is your-- in your opinion, . Jonas? What do you think?

I Agree. OK. Paola asks, "Thank you, Jonas and Adrian. Fantastic work. It is clear that Botometer is not reliable to identify bots, but what other methods you recommend? I agree that bots are not the major problem, but they still pollute the pollute the public sphere."

And I think the quick answer to that is to kind like get to other questions as well is do several methods at the same time. Always manually validate, which is kind of like, you know, annoying. You kind of like do automation to get away from this because it's just like time-intensive, but, you know, like you have to do this.

And in any instance, I personally like network method detection, like Network Detection, but these are also very time-intensive. And there, you do have to kind of like manually validate. And so what do you think Adrian?

Yeah, I totally agree. There are methods to identify them, but maybe clearly about this really means, like you need to look into the data and do some data crunching. And look at it from different perspectives, but definitely not use a classifier in that process.

OK, Yen Ting from NTU, Adrian, asks, "Have you already figured out that-- so since you have already figured out that the diagnostic ability of Botometer will differ because of languages. But do you think that different social media would also interfere with the precision of Botometer, although all in English?"

Like, we use the same classifier for a different social media platform. Yeah, definitely. If the same-- let's assume there's another social media platform that has exactly the same affordances, has more or less, right. What you get back is exactly the same you have something like shares, which is a retweet. You have replies. You have mentions. All the same exists.

I would say, of course, there are always cultural differences. We can even stay within Twitter, right? Or we can move to another platform where it may be even more obvious, not something like Instagram.

In some countries, for example, the kind of Comments section is extremely important, right, and a lot of things are happening there. And in other cultures, maybe the Comments section is not so important. So if you create a classifier, and in one culture, where the comments are not so important you classify users that post a lot of comments as kind of like bots because there is automation, and then you use that classifier in the other culture. It would probably kind of like identify a lot of your users that really write comments as bots.

So we also need to be aware, not only about like it's a different platform with different affordances, we also should consider within one platform, of course, there are different cultural spheres in which social media are used differently.

Awesome, thank you. And so the final question of the day, or the night, for you, it is, what is the ideal way to create gold standard data sets for bot research? Is it possible for humans to classify bots in the wild?

I think there are two parts. Gold-standard within communication, science, gold standard within computer science, so let me, because we don't have enough time, let me talk about the one in Communication Science. So like I say with the sentiment analysis, we assume we as human coders can identify the sentiment if we read the text. But we know, also, from content analysis, sometimes there's a lot of ambiguity. It's not that simple. Right.

If we use the same for bots, we have to assume that we, as humans, can recognize bots. So the gold standard, in that sense, is like you, as a human, take a sample from the population and

manually classify these accounts, where you can even give a score, right? For the bottiness, you could check that. Right?

And then you compare the values you get with the values that the automatic classifier gets. So from a communications science perspective, of course, the gold standard. But I would say, or I think also most of my colleagues would agree with this, I would read as human coders are seeing. And then you can even go further.

Actually, in communication science, we say, like, it's not enough that just one coder, especially with sentiment, that would say, it's not just enough that I check it because if we just compare sentiment analysis between what Jonas sees in a tweet and what I see in the tweet, even if we are from the same culture, probably there is not like 100% overlap. Right?

You would maybe even have to add more than one human coder, and then you can compare the kind of like labels that we give as human coders. This is the gold standard, obviously, with the-- what the classifier gives.

So, in our case, we can test, in this case, here, a Botometer because we took all these accounts. Right? It's very clear, these accounts we have selected, tells you the bots or humans. But I would say, in the general Twitter population, there are also actually probably bots that are not just labeled as bots. And then you need, actually, a human, who first has to identify them.

And that's like manual coding. So the gold standard, again, to answer this question, in my opinion, it's like the human coding, maybe even more than one human coder. What do you think, Jonas?

I think that sounds about right. And, due to time, I will just refer to Adrian's perfect answer. And, with that, I think we all can wrap up this event. Thank you so much for attending.

I hope, if you've got questions, please feel free to reach out to us here in Twitter.

I think the chat is still open.

Yeah. Yes, I believe we are. Twitter, we are, you can also find us on our institutions' websites. I hope you all stay safe and healthy. And we'll see each other with the next virtual event. Thank you so much.

Thanks for joining. Bye bye.

Bye.