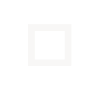
Transcript

March 25, 2024, 11:00AM

  
Fabio Pierazzi 0:03  
You're in charge.

  
Changjoon Park started transcription

  
Fabio Pierazzi 0:06  
We don't have this small teams right.  
I don't remember if you remember if I created a small challenge, a small channel on teams with the four of you, with all of us.  
I didn't.  
Uh, because, I mean, it's just that for a quick short, maybe I'll create after this meeting as well.  
Uh, it's just, uh, it's would be just a small teams chat where it's going to be all of us.  
Uh, it's just for convenience.  
You can ask questions for to the others, but if you have a very very quick simple question, you can also look a message there like of course for more complex stuff it's better just to wait for this new games for or if I may use the chat as well to say, hey, I'm uh, I don't know.  
It happens that sometimes the train doesn't work or like there's a.  
If I have any predicament for which I can't join the meeting last May not having just the chat, there is a bit venient and finding secret emails, some remembering so also for last three of communication sector that I may use that chart as well.  
But I shared in the chat of this meeting just a link to the assessment criteria of the report.  
It's a little think about that, but.  
It's I always because I also have passed already septations finishing now and I think it's important from the start that you get familiar with the sections that you're expected to do.  
And the and.  
They're all sections except general scholarship in the link I gave you.  
Now umm.  
Sorry, I I think about sections because I think about research papers.  
What I'd like chapters that would be in your final report, and because when it comes at the very end of the project, it comes to marking, this is the breakdown of the marking.  
Basically, we have to mark each chapter separately and this is this tells you the weight of each chapter as well and the criteria for marking this is very general criteria, so it may not apply directly the project you're doing here.  
Umm, but it's the chapters need to be those.  
Basically, the name can be a little bit different, but you will have an introduction setting the context and that will give you of course more uh advice was shown what should be there, but it's usually a couple of pages.  
The researcher review, which is also what you're informally doing now and will be part of the credibility reporter.  
And here I would say minor description.  
Sorry, there will be a minor distinction between background and literature review.  
Doubting the same chapter usually, but background is just textbook knowledge like I don't know you just Rundle Forest.  
This is how the Rundle Forest works knowledge.  
Which are should agree studies like, OK, you're doing, you're trying to solve a certain problem.  
What other research approaches have been published and tried to solve similar problems and how you position with that?  
And then there's a chapters about the design, let's say objective specification design.  
Because I know that some of you must scarifications as well.  
About what design really means and in uh, comparison with implementation.  
So in relation to implementation, cause design and implementation are two separate chapters in the design.  
Ideally, you shouldn't choose such using any line of code.  
You shouldn't be citing any specific library in the designer.  
It just.  
Uh, you can see there's the abstract view of your project.  
How you planning to solve the objectives?  
Having like flow charts of the pipeline that you want to use with moderated system, but ideally people can implement that in ideally even different programming language.  
So so you need to write in just that abstract level.  
For example, random forest is just the general algorithm, it's abstract, and then you have different libraries that implement from node.  
For us, with different tradeoffs and different parameters.  
So in the design, just think about at the abstract level, how you're planning to solve your project and the design of your system.  
In this case it will be more of an evaluation pipeline, right?  
You mostly see these on papers in the other sections, like a diagram showcasing the major steps you're taking, and maybe take another set, or you filter it and split it.  
And basically it represent a clear visual representation.  
So how you do that with if course the text to all discussion.  
In implementation chapter you talk about how you decided to implement it.  
Which libraries to be used to implement it the the how did you structure the coding case?  
How uh.  
If there was anything particularly challenging to implement, or if there was any particular, I mean maybe there was something that was like doing I computational cost.  
It took a lot of time, so if you had to come up with a solution to solve it, you can also discuss that into the implementation chapter.  
Sometimes it depends, right?  
So if it's a more of a theoretical problem, you can talk about that in the design chapter.  
But if it's a library problem, OK, mutation of the library specifically and not of idea of the approach.  
In general, you should talk that well that the implementation chapter and then.  
You will have the chapter on results and now as you said, they always show, which again is one of the chunky ones where you discuss, uh, the results of the experiments that ideally where described in the design chapter.  
And then there was the legal, societal, and that call.  
Sorry, legal, social, legal, social, ethical and professional issues chapter.  
Which is can be just the page or true.  
It just shows like you need to reflect on implications of how you connected the researcher and if the research can be misused, the precautions should maintain, take and blah blah blah right?  
I mean blah blah is meaning, like it doesn't have to be super long.  
It needs to be there and the little bit confusingly in this structure, general scholarship, which is, which is not really a chapter, but it's more of a separate mark for uh.  
The general's collar ship in how you have written your report that have you cited the papers correctly?  
Is it well formatted?  
It do you write sentences structured well so the just more of a general?  
It's like general scholarship is not cheap there somewhere.  
All I'll show you as well.  
So, or at least that will clarify, but I want you to say right now, even though it's not this immediately your problem so that you start thinking about the project in this terms.  
And I'm saying this because all of students who realize it later on also because the structure for the preliminary report that you will have to write is just the partial part of this and.  
So yeah, just I would advise you to, you know and then just refresh this stuff, maybe writing bullets, but what you would write in each section because it will also give you a little bit of an idea of let's say.  
What's your?  
How?  
How you're doing in the project as well?  
So I mean ideally of course you will see that each chapter is different weights as well.  
So I mean, definitely understand that the results analysis are worth 20%, for example, so.  
Felt 0 results.  
Of course, that's a problem as well, right? So.  
If you want to balance.  
To to understand how your how you're doing right?  
Or maybe you don't have anything at all to writing implementation chapter and you need to reflect on.  
OK, but what can I do to put something in there?  
Should I structure the going to be better?  
I don't remember if this was the case for the Ms C projects as well, because for the bachelor at least it could be that the the documentation that you have in the Keats page is a bit more focused towards software engineering projects.  
In this case, let's say the according part remains important, but it's.  
But there'll should design and the 90s part, which is, in a sense more important.  
So it's a little bit more research focused than provide focused, right?  
So maybe I will try also to scratch now something here that you can follow as well and they don't remember if I told all of you.  
But in the.  
If you tell I also put a dissertation sample.  
Umm, it's about shortly statistical sample, but it's an excellent sample.  
It's scored over 90% of don't get scared too much about this specific research because it was also follow up of.  
So it happened that the student also did the curve with me before.  
Four.  
So he had a little bit of experience already and on the topic.  
Umm, so the project may look very advanced, but that's also because in a sense it's kind of a follow up to another project so don't necessary.  
Don't feel intimidated by the fact that maybe the topic was very, very.  
Complex.  
Maybe have you thought experiments?  
But it's anyway good, I think to keep as like of.  
If not, yeah, I think at the end of the day, again, it's a very, very good project.  
It's scored over 90%, so it's good to keep keep the best as a reference as as sounds perfect either anyway, but overall it was a very good project report as well.  
So you're going to search for that.  
I think this structure should be basically the same.  
Maybe.  
But in any case, refer to this system criterion kids.  
I OK.  
I just wanted to.  
Uh.  
Sorry for this, I know it's certainly advance.  
It's not the problem we have right now, but I think it's important to start at this from Hartley on and and by the way, there's a 10% oral presentation as well.  
Let's say 10% is like 10%, so don't mark.  
You will have to carry a video presenting your results and that's why I was.  
So we'll give you feedback sometimes on this that's presented here.  
So if it says any feedback I give you this nicely present here.  
Again, take it in mind.  
Typically, keep in mind for the final size it would do.  
It's going to be a bit different, of course, because now you may be have a slide with questions, but I think I'm talking, I will advise you based on like if there's too much tax, too, less tax, too much images, too, less images, stuff like this can keep this in mind or just feedback for the final presentation you're going to do.  
OK.  
So that's really too much.  
Is there anyone that needs to leave early?  
If I can ask her.  
So if do you want to be specific order?  
Otherwise, I'll ask her that maybe to go first, though.  
OK, which is fine.  
I was gonna get.  
You should just share that so there's a the top right, there's a share.  
Ohh OK so I just figured that.

  
Yichen Zhou 12:54  
Sorry, I need to go to the toilet.

  
Fabio Pierazzi 12:58  
Fine, thanks.  
OK.  
Because they just get used to.  
Team before that have like.  
From four point itself, like plug in that you can use.  
But now I think people with ohh the new one.  
Uh, but is OK.  
I can't.  
I mean, yeah, you cannot shop floor if you in the share if you upload.  
Let me just ask if you want someone else to go first or if it's if you feel bloated and if you're sharing now.  
Maybe someone did because I have to turn it off and ohk OK also be June, do you want to go?

  
Jun Shi 13:58  
Yeah.  
Yeah.  
OK, I will go first. Uh.  
So.

  
Fabio Pierazzi 14:06  
And again, if you're typing, it's taking some notes.

  
Jun Shi 14:07  
That let me share my screen.  
Standard. Ohh.  
So can you see it?

  
Fabio Pierazzi 14:23  
Yes, yes.

  
Jun Shi 14:24  
OK, so how do everyone?  
I'm gonna do presentation for my variants, used and requested features.  
So firstly, the approved figure uh reviews and shows the objectives of the three phases talked about in the previous meetings.  
So and then the following figures paragraph shows the basic experiment.  
So faces and tasks.  
Since I I did some umm as you can see over here Sir, I did some processing of the data size and so I will here you can find I created a filter data side that only included includes the used and it has the features inside and then comes to the preparation of the model.  
Yeah, most of the models and experiments I used, I tried some machine learning models such as SVM, random forest and basin.  
I haven't tried the deep learning I want to, maybe I probably do it later, but I just.  
I also got some outputs from these three machine learning model.  
I'm gonna display it later and after the model using, I'm gonna explain the and I'm going to explaining and their size so.  
So I I'm gonna use different shop expander for Tristan for different model and once a certain amount of the results have been produced the work can be begin on describing the results of the project.  
So which usually are description of the data size presentation outputs of the classification results and maybe some and also the most important thing is the graphical uh outpost out the expanding results.  
And here is the the stages marked in yellow as the part where I'm currently experimenting and so as and I county and only tried the.  
Or model with default parameters.  
I didn't do any.  
Do anything with the fine tuning training, but I will do it so I hopefully I can slide the best one as the solution.  
I mean here the solution I mean I I will compare the results that's from different fun tuning model and then let's discuss which one maybe is better and also the model, the training model parts gonna with four data sets and failed data size.  
And then I also cite some model evaluation such as accuracy score for every score.  
I just print some classification classification reports and then I display the RAC Curve and official matrix hoping this.  
Evaluating results can predict and estimate the expected performance of the used data side.  
The right stages are where I'm about to do that.  
I'm I'm experiments are facing some difficulties when trying shop so so I will uh, maybe a set some maybe my question.  
So on then there's a workflow for the methodology or the experimental design.  
So for data analysis and visualization, gonna check the quantities statistically.  
It's besty statistics and some data distribution distribution.  
Check and then going through dataset filtering and for I splitting the data sides and then I will proposing I will propose the mode machinery model with two data sites and for and after finish the this the 1st 4 stages.  
And because I'm also wanna trying to.  
Uh, do some challenges.  
Challenges such as concept drift or the shortcut learning parts, so I will response to the change part problem with metadata and by repeating the uh, experimental method or the stage above.  
Here I display some common currents.  
Works with results.  
But for says this slide is for the full data size and but for the model I didn't, I didn't do any fun tuning, so both from the default parameters and as you can see over here, I think the best results shown over here is the is the SVM they gave the good if each score and.  
And the printed.  
I think the good RC and RC curve.  
Umm.  
Then as this one is for only filtering for the permissions features, that's still ask me.  
I am give us the good results but I will keep trying and do fine tuning with that and I I'm also gonna compare all of the evaluation results and let's see what's happened.  
So here is the some challenge stages that I'm faced.  
I'm facing so the first one is the technical issues of applying SHAP with a large data size.  
So umm the the.  
But I mean, that's the policy might work.  
My clothes done.  
Work so I'm.  
I'm still I'm.  
I'm still looking at the official documents and how to adjust FISA parameters so and I'm.  
I'm also trying to understand the email.  
That's Professor sent to us.  
So the second question is because I we I have I created the filtered data size only include the.  
Permissions features.  
So I want to know how to compare its experimental results with the full full data set.  
I mean, sorry there's some mistake and how to describe why there is it is effective in detecting more well and the third part is the structure of the uh for any preliminary reports cause.  
We need to submit the reports on I I think that's the end of the April.  
So I hope to know how how to manage the structure.  
So I think that's all.

  
Fabio Pierazzi 22:44  
Yes.

  
Jun Shi 22:44  
Thank you.

  
Fabio Pierazzi 22:46  
OK.  
Thank you very much.  
First of all, I think I think it's very good that you're you have results for the for the feature set some classifiers and that your knowledge it's without tuning yet.  
I do have some comment, maybe I'll say just the main ones here.  
Also, to answer your question and then I'll message you the rest, because I anyway took some notes.

  
Jun Shi 23:12  
OK.

  
Fabio Pierazzi 23:15  
For the preliminary report, the structure is in the Keith page.  
There's a [MSC preliminary project Report tab, and the structure and report format is in the label, so it's in the description of the kids tab before the submission link you just sharing a quick screenshot and I guess it also anyway follows the description that I gave at the very beginning, right.

  
Jun Shi 23:43  
OK.  
OK.

  
Fabio Pierazzi 23:48  
I mean, of course it's just part of what it was out of any beginning, but.

  
Jun Shi 23:48  
Thank you.

  
Fabio Pierazzi 23:54  
Uh so 4.  
It's good that you also had procedure that recall curves.  
Or since the data set is highly imbalanced, though for the AUC under the rock curve, you should zoom in on the X axis for low false positives, so the X axis now goes from 0% to 100%, or from zero to 1.

  
Jun Shi 24:21  
Uh.

  
Fabio Pierazzi 24:33  
Should probably focus on ohh from zero to 0.1 which is 10%.

  
Jun Shi 24:42  
OK.

  
Fabio Pierazzi 24:43  
And if you want to read more about why, read about the base rate fallacy in.  
For the tuning, keep in mind for tuning the algorithms beyond their default parameters.  
Keep in mind that ideally you should divide that data setting three parts, training, motivation and testing.  
And you should find tune by using all the training and validation.  
So otherwise, if you just keep it training in the testing center, you find tune, you can play with numbers a bit and you will get a inflated performance.  
So just keep three separate sets, but and then the first stages it's fine if you just separate them randomly, like 80.  
Sorry, 60% training, 20% mediation, 20% testing.  
But you can also do differently.  
Some people may also do 8010 time and for sharper, I think it central swimmer write about sharp.  
I think that the main problem you may be encountering is because of the perturbation distribution.  
Uh, I did send an email to you.  
Alright, OK.

  
Jun Shi 26:03  
Yeah, yeah.

  
Fabio Pierazzi 26:05  
And but again.  
Let's say uh.  
Try to really read the parameters as well.  
I think the main issue may be about the the this distribution.  
Now I will also create a chat, but you're free to talk to each other about how to solve this as well.  
I think it's really a parameter problem.  
Like if you shop all of the boxes will just be infinitely low.  
Uh, one thing you can try is those who just uh, subsampling first or even do some toy examples first so that you just see even how the computation time scales up?  
I guess that then we'll show will show off and out of time varies because for me, a lot of time is a multiple days, but I understand that if you're running locally on your laptop for you a lot of time can be one hour.  
Uh, swing.  
That device is also for avoiding to recompute everything all the time.  
If you use the pickle format, you may also save temporary training models so you don't have to do it all the time from scratch.  
So look into like saving as peoples intermediate results and models so that you in your in your call time when you restart the code the model is already trained so it will be a bit faster as well.  
But again and then I'll I always send you some details, some other comments in E text, but.  
Do you have other questions so far?  
I mean did did I answer most of the things or?

  
Jun Shi 27:50  
Yes.  
Yes, thank you.

  
Fabio Pierazzi 27:54  
Cool.  
So if you think you're ready.  
So we said go next.  
Sure.  
Right.  
OK.  
Yeah, just make sure that, OK?  
Uh, OK.  
Hello, my name is Sid and today I will represent my progress so far on the project.  
So so far I try to understand the data set that we have.  
We use the data set and rezone dataset and because it's all under zone, that is, it is almost 50 terabyte.  
There is sample of it that we use to have our experiment, so I just take the data center to have some virtualization to understand how this data is it look like and after that I have I do a classification for whole feature set to understand all the features.  
So with it looked like wipline and each time I eliminate one of the features feature space and I start with the Roc, care to see the performance generally overall and I see almost all of them have nearly same result 98 to 94.  
So I said that, OK, I will go more deep and to understand the dataset more and.  
To understand this I use SHAP to have explanation.  
So I take the to 20 of the feature space.  
If you see on the left I I use I remove the total, I mean the whole name of the feature is the feature and just put the briefing.  
Yeah, just put the prefix too to see it more clearly.  
So I I say from the top 20, I see a lot of related to your.  
So I just say why you aren't.  
So I come back to the data set to compare, so when I just go and see if the oral, I see that most of the data sets have sampled related to the oral.  
So this little bit makes sense for me and I I, I and one question was in my mind why I didn't have from the top 20 any feature based related to mine like activities and receivers and and others.  
So I just go further.  
More umm and do small experiment before that.  
So late to the the the Daniel they and they literally they take small malware time family dataset and they compare all of these family and they see all of them have permission features set.  
So overall, they decide or conclude that all features that is important so.  
After I eliminate every sample uh to make He come conversation, I mean comparing between all of these I have here table so I'll show the precision, recall and F1 the following scope.  
Sorry and I see that all feature and for example the second one API calls have almost same values.  
So if I would eliminate that be called, this will not affect the other features except and will give me same result.  
So from the this experiment we can see that we can eliminate some feature space without a fixed the performance and we can see I just take a sample of all IB calls here and come back to the data set to see the virtualization for it.  
And I see that they they have like good proportion for the dataset.  
So this and Phasis my my result and and sorry comparing to the activity and services and services and receivers my feature I see that they was not on the top 20 and the sub explanation so far and I see.  
So I come back and try to understand the dataset why activity and services and receivers feature was not from the top 20 even when they was go up.  
So I try to understand the data set huge Willy and I see that last two years the feature set go up as a malware and there is a huge portion of it.  
And this is not reflected on the training itself.  
So it was like 1 big question for me why the data?  
Why the train?  
The training model not take this feature even when it was in the last year.  
As we can see, all of these uh, my feature was increased on the last two years generally because I just sorry can can can you go back to like the histogram?  
Yeah, we're second tribulation here.  
What do you mean?  
Sorry, when you say contribution here, what do you mean for the whole data set?  
For example, in 2018 3037% of the malware was activities and services.  
From the date has been set for the sample contain the activities.  
Yeah.  
So you I try to understand the.  
No, no, I mean, yeah.  
But so first of all, also the equation is like is it?  
So the red column is really 37% out of my.  
We're only yes for its in the data set.  
It's out of nowhere.  
Ohh yes, and the contribution is like if they have at least one of these features.  
Yes, OK.  
In general, if they're any of these features that to one yes, OK, if they continue, it's OK.  
It's a bit weird.  
Are you using the tank case at?  
There's more product they're used for the holidays.  
OK, you know this is a bit old the right because I mean I wouldn't mind activity can be even just a window.  
So I would imagine that the I mean.  
99% of the absolute of activities. Uh.  
But it could be that.  
But anyway sorry.  
OK. I understand.  
Yeah, I understand.  
It's all you're speaking that relative just the login screen.  
So any app install have at least an activity usually, yeah, but uh, it could be something wrong with the my.  
So there's something wrong other other with your visualization.  
Yes, for could be.  
Also, I mean there could be a little bit of like can imagine that in the in the data set that could be a little bit of.  
It could be that we don't mistakes into that satisfying the feature extraction.  
I mean, but I wouldn't expect it to be so major.  
Umm, all my clothes look like 2 for my first.  
What is this in or from the dataset as that's struction of the feature space?  
Maybe the sample of the apps that are taking I didn't or I think mine.  
I have to take.  
Yeah, maybe, but I think we'll finish. Sorry.  
Like, I'm sorry for interrupting.  
I just wanted to be sure.  
So, OK, OK.  
So I just try to understand my features space and I see that it's go up last two years and what conclusion?  
Ohh the experiment I I did show that some feature space can be eliminated without the fix that we forms of classifier and the best I know so far more Research linked, not linked trend of software development with the malicious app.  
So we not have enough study study related to the feature space itself.  
Maybe it's trained on software, or maybe it's ohh abuse from the from the attacker.  
Scenery, umm.  
And that's all.  
Thank you.  
OK.  
Thanks and.  
Or as a rule, Ferrari will never feel.  
I mean, I know that you don't know.  
Maybe so.  
It's fine, but never finish with it.  
Thank you.  
Slide or Q slide.  
In general in life, when you finish on this slide, it's better.  
It's better if it's like the conclusions lines or something.  
OK, so that a if people have questions they and even if they lost track, they lost the live is to provide context for them and you leave it on and there will be a cure in any session maybe I mean not in the project per se but just in life.  
I know that people still do it sometimes.  
I mean, it's just a umm, so again good for having results.  
Though I do have some technical comments, the whenever your shoulders lights are generally good, but there are, I think two major things to improve versus the definition of the.  
So for example, contribution is not defined at all.  
Hmm.  
Why ask the question should you should find these metrics as well when you presented in this slides and it's good that you keep this slides relatively empty.  
But when you have clothes, it's would be ideal to have couple of mullets.  
Of course you need to explain my voice, so it's fine, but the couple of bullets with every key highlights, or even just the question or trying to answer the title, maybe if this slide.  
Practically.  
So if you go back to this line with the Orcs, the course. Ohh.  
Yes.  
So I think this is again, as I was saying to Jun before base rate fallacy, they looked very similar but you have to zoom in, OK you have to zoom in because anything up imagine like so I'm gonna use the mouse here.  
So people are not going to see, but anything above this point, this system is just completely unusable.  
OK, so you have to basically zoom in.  
When I say zoom in with something, just change the next axis to look at the first part and then ideally you'll show.  
Maybe you can look at.  
I don't know.  
Even maybe or should the Y axis.  
You can probably filter from 80% maybe because anything below 80% is not really good.  
And then you will see that there is a difference, yes, as you will observe here, which was considered a confirmed also with some other students in the past the URLs are performing actually way worse.  
This is not too much reflected if you go next to the F1 score comparisons.  
Yes, I think so.  
For which model are you using here CBC and so the linear SVM yes, with C equal I mean we all the stuff I can check But it's I think it's.  
I think it's from the from this simple example in the GitHub and OK OK so like 60 because I would imagine so is this a random split or how did you do the separation of training and testing?  
To be honest, I didn't focus on find way to to working.  
Ohh, but it's looked like 6030 last time I see.  
Maybe I can check for if I will.  
No, no, no, it's fine.  
But it's run those split to doing.  
I don't remember.  
I mean, I don't remember the code that.  
Yeah.  
Yeah.  
In the guest hub, I think 1613 I don't know because I would imagine it should be a little bit higher.  
Maybe double check?  
I mean, if I remember correctly, it should have been 90.  
Troopers and maybe OK now with all the features, but but again, maybe I remember wrong.  
You may depend a little bit was shown how much you're sprinting as well.  
OK, it could also depend that if you get the lucky splint or not.  
So you can also set the random state with certain value if you want to give experiments, but anyway the blue so but I I don't know so this I would double check a little bit maybe because no OK for the URLs it's much worse which is fine.  
That's what I know.  
I just have for the others.  
I think it's fairly similar.  
Uh.  
Maybe.  
Anyway, what I mean is just maybe double check that it's done right for the contributions of features so.  
If you talk to anyone outside of this, I mean, if you talk to another researcher and say contribution of features, especially with the SVM that we think about the weights.  
So and which is something you may also want to look at versus the presents from sensing that auto said, I think please maybe that will check your histograms because I don't know.  
I feel there might be something off and so I would imagine that the percentages were.  
The feature appears solely the percentage of samples with them malware where it appears at least once should be higher.  
I feel especially for the activities, OK, it could be there's a problem with the feature extraction because I also have another version of the data setting case, but I don't think it should be that major.  
So maybe just double check it's computing correctly.  
OK, I'll shift the code.  
I I wouldn't say it's the selling.  
So it's good to double check or for the feature extraction, but from the contribution perspective I don't think it's super super important in the sense that.  
What is it?  
Maybe that you were presenting like out of the features in the malware 32% or activities.  
This is that what you were trying to represent?  
No, it's like the following. Appears.  
Yeah, generally, yeah.  
Appear.  
Generally the all the data set.  
If so, yeah, it's not be quiet, but anyway appear sharp since the last set.  
It's not the silly telling you much, and the reason is.  
In the so it depends that you want to have a similar feature between all the malware.  
So in in the classes so malware need to have similar features on itself and goodware need to have similar features values one themselves.  
In this case is presences.  
Uh, but then you machine learning tries to maximize.  
Uh that each like good learning for very different reasons from malware.  
OK, so the fact that you have a lot of activities doesn't necessarily mean much if if, let's say if inactivity like the login activity set login activity would be that we N with the logging screen in another app.  
OK, it could be that it appears a lot in both malware and goodware.  
If it appears, let's say that that they can.  
Extreme scenario.  
Let's say that all the apps have been activity.  
All the logging for login if all the apps have it, it doesn't matter at all for detection, but it's just ignore what I classifier so you see if you just compute the.  
Quantity of time should appear, so it's not necessarily indicative of how relevant it will be for detection.  
What you can do is look at the SVM feature weights.  
I think I added a script, but anyway the classifier as a parameter called the coefficient like underscore.  
Q.  
If OK.  
Uh, which is an array of arrays.  
So you have to do quite of the role to get the array of weights.  
And since you're sitting across the fire, where it will tell you the contribution of issue which tell you the contribution of each individual feature, so it's closer to what I think you were trying to do means and then you can group the weights by category and maybe even do a box scroll to retribution of weights.  
OK.  
And you can maybe compare these as well.  
OK.  
And it's just but.  
But I know, OK.  
It's just it was just a comment on uh, but yes, generally it's good, let's say, just be careful about uh, there are curves.  
Zoom in on the low for responsive rate region and hydro positive rate region so she the using procedure called Jun did cause procedure recall curves are less sensitive to the other setting balance OK and check again this contribution thing.  
Yes.  
What did you guys say it already?  
I don't know, but yes, just be careful in the experiments that you do and double check will show that this table over here.  
But I mean it looks good as a direction, so it's good that it got some results already.  
So yeah, I have one question.  
Sorry about recall, when I eliminate the last one for the URL.  
Yeah, example there is precision once almost similar the as the previous one and force was from 84 to 70, but the recall was dropped down to 0.5, so that's.  
Is it like one question for me why that he called was slow and the decision was high and for a performance call was I mean you may need to be we require a little bit more investigation, but I would imagine that.  
But this accurate to use this where the way are you retraining a classifier for everyone of these roles yes.  
So it could be that the.  
So it could be that.  
The URLs.  
So basically, what do you what this means is that but it it, it made me some hour based on your own features, but it will not generate.  
OK.  
I mean still any well satisfied, but let's say the Hyde procedure in tasks with that it will not generate many false positives.  
Yeah.  
So what I imagine is what I imagine is that there are a lot of, so there's an area with a lot of future and malware with similar URL features.  
Sorry, but this after any minute you aren't, so I just have no.  
Ohh 33.  
No, wait.  
OK, so double check it on the thing actually, because I know that with URLs only it should be.  
OK, I see what you mean this.  
OK, I'll try.  
What you mean because you've done a different thing?  
So the the.  
That's how the meeting right?  
Also, I have on my I just only my feature and then recall was down.  
So there is false negative or not.  
It's like this.  
Is that the classifier have many for each other.  
When you have a drop in accuracy because you remove certain feature, you should was very important for the specification as well.  
So in a sense, without the URL.  
So it seems it it is harder.  
Or the classifier.  
So there must be a region where there's a lot of confusion between good and malware in a sense, and the fact that the precision is high.  
Yes, it means that you just change because that's something you can change as the threshold.  
So you can also change the threshold to increase the recall, but then it will result in a ohh it would result in lower so there's a procedure called tradeoff always if you check imagine that detection threshold you're gonna say it's malware.  
If the probability is the 50%.  
Or you can change that if you want to be very, very confident you can say it's malware.  
If it's over 90% accuracy in the like if the classifier gives you auto probability, you can say I want to just be sure it's really ours, so I'm just classifying as malware.  
Anything with confidence higher than 90%?  
In this case you will have lust for supposedly, but a lot of statives.  
OK, because we're missing a lot of stuff.  
And I don't know.  
I guess we're moving.  
The URLs will increase the confusion, but I would try to replicate.  
I would try to redo this experiment a bit or like double check it's done correctly, because I don't know.  
I would have imagined that also, for example, we're moving API calls through them and they have bigger impact.  
And yeah, I'll so it's not clear to me why you do it with activities as well, because if you know it with activities should be also quite travel for massification anyway.  
So if you remove activities and we'll expect that as a drop in performance.  
OK, I didn't.  
So you gonna says this is the way.  
This is one way to look at feature importance as well in directly, so it's called.  
It's based on an application.  
Study your own the feature set.  
Actually, it's called descriptive accuracy.  
You're all the feature set and if the detection accuracy drops a lot, it means that that feature set was important.  
OK.  
Uh, so we return to redesigned experiment a bit, but yes, it makes sense in general.  
And of course introduce show feature categories as well.  
And sorry, I'm conscious of time that are really running over and I think I'm maybe asking the Cheng to go next.  
Let's see.

  
Yichen Zhou 51:19  
OK.

  
Fabio Pierazzi 51:24  
Particular remember the last time I took it last and we run.

  
Yichen Zhou 51:40  
I'm not sharing my slides right.

  
Fabio Pierazzi 51:43  
Yes, but I don't.  
I see them in PowerPoint, not in the full now.  
I see full screen, yes.

  
Yichen Zhou 51:49  
OK, so the research question is if in the malware analysis project, rather than using all types of features, the machine learning algorithm only uses one category of the features for the analysis, which is the URL, will this affect its performance and accuracy?  
Can explanation methods help us to understand how to improve the design of the detection algorithms?  
So I'm using supervised learning algorithms.  
The algorithms that have a labels in the data set to train the models used to make the classification decisions or to predict outcomes.  
The labels tell machine learning models what the correct decisions are so they can learn if the decisions they have made are correct.  
The data set is separated into training and testing sets in the training sets.  
Uh, it stores.  
What the machine learning models used to learn and in the testing set it stores the data set used to test the performance of the machine learning model after period of training.  
I have used the three kinds of machine learning algorithms.  
The first one is logistic regression is a classification algorithm that estimates the probability of event occurring such as voted or didn't vote based on a given data set of independent variables.  
Naive base.  
It's a it's a family of generative learning algorithms.  
It's used for classification tasks too, and more specifically I'm using multinomial and B and there is a algorithm linear as we see, which is the default of the in the code that is already provided in the code.  
So the goal is to test the model which only uses the feature for the categories are URL.  
So we need to understand that what is in the data set accident DOT Jason contains the list of some dictionaries.  
Each dictionary represents a software.  
Each dictionary contains the feature of the software to be analyzed.  
The keys of the elements in this dictionary shows that feature categories and what we need are the elements having the keys containing the keyword you URL.  
Why dot Jason contains a list which shows the labels for the software.  
The software can be either malware represented by the integer one, or benign software represented by the integer 0.  
Each element in the list corresponds to one dictionary element in the list of X dot Jason.  
So uh, methodologies and is a metadata and the other files are are the reduced data set.  
I didn't use it in the experiment.  
So we need to get the test result for URL features only.  
I use the new data sets that only contains the URL features.  
So here is how I provided how I prepare for the data set.  
I loaded the data sets extort Jason why dot Jason and store the data in variables.  
I create new datasets temp X dot Jason and temp Y dot Jason.  
I declared a new variable named temp Data X and the temp data Y which store the data required for the new datasets according to what features the software have.  
If the software has features containing the keyword URL, store the features of the software having the keyword URL in the variable temp data X and the store the corresponding labels for the software in temp data.  
Why so?  
I created two data set files, open the data set files with writing permission for putting data in them and finally store the data in temp data X and the temp data Y in the data set files.  
Now the data meets the requirements of processing the features.  
Only then I continue with the analysis on the new data set files.  
Uh shop will be tried to understand how to improve the design of detection algorithms.  
It's a way to explain the output of any machine learning model.  
It uses a game theoretic approach that measures each player's contribution to the final outcome.

  
Fabio Pierazzi 56:29  
It's.

  
Yichen Zhou 56:35  
Uh, the requirement is to get the statistics of the data set and the test results.  
Compare different classification algorithms.  
Compare the results of the tests with all features and test with only the URL features.  
I then I can get the conclusion which is whether using a single feature URL as it is as effective as using all the features.

  
Fabio Pierazzi 56:50  
Sleep.  
It's it's, it's.

  
Yichen Zhou 57:05  
Finally I researched on what has caused the experiment results.

  
Fabio Pierazzi 57:10  
It's.

  
Yichen Zhou 57:12  
The methodology I used, including test the machine learning algorithms, show the statistics and compare the results by F1 score, accuracy, precision, recall, confusion matrix and the analysis with sharp which will be done in the which will be done in the future.  
Uh, the confusion matrix shows false positives, false negatives for true positives and true negatives in the in the diagram.  
The confusion matrix can be used to calculate the performance and accuracy of the model.  
The accuracy is the proportion of the correct classifications in the model, so precision is the proportion of positive class predictions that can set actually belong to the class.  
In the question, the recall represents the percentage of class instances detected by a model.  
The F1 score includes is the harmonic mean of precision and recall.  
It shows the performance.  
So here is some confusion matrix of that of these algorithms with and without.  
Uh, are the other features?

  
Fabio Pierazzi 58:29  
It's.

  
Yichen Zhou 58:29  
This is a statistics of the F1 score, so the F1 score.  
I found that's F1 score decreases if if I only use the URL feature and also the accuracy also decreases slightly but not very much.  
Here's some other statistics, and the conclusion is that only using the URL features will significantly affect the performance of the MLA.  
That can identify malware if considering the balance between the procedure and the recall, the accuracy will also be affected by the effect is not very significant.  
So my question is, I have something to ask about the reference style and I don't.  
I'm not sure what to add to the design part.

  
Fabio Pierazzi 59:20  
OK.  
Thanks for umm.  
Thanks for the presentation.  
So I do have some general feedback and I'm conscious also of time because stay and I hope I also Changjoon can stay.  
Uh, well, if anyone of such to leave earlier, it's fine.  
Let's try to wrap up.  
I will send you some more detailed feedback afterwards, but.  
What you described in the designs lies.  
I feel this implementation stuff, at least in terms of level of detail.  
You talk about Jason files, you talk about dictionary structures.  
You've talked about detail.  
Basically, you're doing a sort of high level description of the code.  
What you need to do in design is to create a higher level abstraction of this.  
You can have like small small circles as the dataset and then you have a filter datasets and then you have to have a few blocks where you just say training.  
Uh crucification explanation and and things like this.  
In general, you can describe also how you're going to separate the dataset without mentioning that it's implementing it.  
I mean it's saved the store.  
As Jason, you're already doesn't matter.  
Xu Bcna sense doesn't matter for the design, I think.  
Uh, I think you do have the content and it's created.  
You have some results.  
Uh, you have to separate this.  
This faster if understand the separation between design and presentation.  
I, as I said to you is like try to rewrite the same implementation without any adding any code details or any dictionary detail.  
Imagine you just have an abstract data set and you want to separate it.  
Conceptually, it features are part of the.

  
Yichen Zhou 1:01:21  
Uh, so the implementation.  
I should not also include that code in the implementation part.

  
Fabio Pierazzi 1:01:30  
Only snippets if they are very relevant, but otherwise the code can go in the appendix, so it doesn't need to be a low level description of the code.  
I think what you presented today is similar to what would be found in an implementation chapter.  
I think that for the design we just have to imagine you have an abstract data set.  
The features are still part of the abstraction because we don't have specific values, so the URL category of features can be mentioned in the design, and then you just maybe also represent visually how you're going to wise so that time and test them, they classify that it shows are also five for the design.  
What you shouldn't talk about is the specific libraries you're using for the implementation of those classifiers, and maybe one thing I want to add is that since you also go to lower results for the URL, you may want to look into.  
Fine tuning the algorithms, like trying to change the hyperparameters and for this, as I was saying before, it's important that you split the data set in three parts, training, validation, testing.  
And you just change the hyperparameters from the validation set.

  
Yichen Zhou 1:03:04  
So I should include the code detail if they are relevant.  
I should only in only if they are relevant.  
I should talk about the libraries I used and I should separate the data into training, validation and testing rather than only training and testing.

  
Fabio Pierazzi 1:03:21  
Yes, in the evening, so.  
The last part can also be in the design chapter, but the rest that you said is in the implementation chapter.

  
Yichen Zhou 1:03:33  
Ohh OK.

  
Fabio Pierazzi 1:03:35  
So the the the separation in training, validation and testing.  
It's also an operational thing that you have to do if you want to try to see.  
So the point is that with the URLs only you've got not so good results, right?  
So you may want to investigate why with chat you may want to.  
Make it work better if you can, and one way would be to just trying to change the parameters of the machine learning algorithms you have.  
What?  
How can you do that?  
You, I mean you basically separate the other setting 3 and you train on the training and you test on the validation set and to change the hyperparameters which the hyperparameters are basically the parameters that you have in the libraries.  
Uh for the classifier like in the Manasquan you have the C parameter and other classifiers who may have other parameters.  
Think I may move on to the to Changjoon, but of course there's also send some feedback afterwards via messenger for some N have taken uh, but yes, I mean, I mean Shao other questions also offer free to just uh, maybe not any email to me, OK.

  
Yichen Zhou 1:05:08  
EI also have a I also have a question about the reference.

  
Fabio Pierazzi 1:05:09  
Chun drum, can you?

  
Yichen Zhou 1:05:12  
Is it that OK to talk it here or send the email?  
I have sent the email I think.

  
Fabio Pierazzi 1:05:16  
Did you?  
So you did, I replied right.

  
Yichen Zhou 1:05:20  
Yes, about that I I had some reference from ibm.ibm.com and they don't have a date.  
So I don't know how to fill in the reference.

  
Fabio Pierazzi 1:05:33  
I mean my email?  
I assure that as well, right?  
Showing citation one and two of the paper I gave you, there is no.

  
Yichen Zhou 1:05:38  
Uh, so they also don't have a date.

  
Fabio Pierazzi 1:05:41  
Yes.

  
Yichen Zhou 1:05:43  
Oh, OK.

  
Fabio Pierazzi 1:05:44  
That's why I that's yes, that's why I I suggested you to just put the last access date that you did.

  
Yichen Zhou 1:05:51  
Thank you.

  
Fabio Pierazzi 1:05:52  
For Dutch page, so not not putting the date of the page is just putting the last access date.  
Your access date like.  
Did you access it in March 2024?  
You can put that datas last accessed March 2024.

  
Yichen Zhou 1:06:09  
OK.

  
Fabio Pierazzi 1:06:12  
The so Changjoon, can you share this slide?

  
Changjoon Park 1:06:18  
Umm yeah.  
I'm before I start, I pulled.  
I apologize for not being not preparing for proprietary slides for this group meetings, but for because of my some the health issues.  
So if it's OK, can I explain my the visual conclusion at visual explanations?  
I can I briefly it's.

  
Fabio Pierazzi 1:06:47  
Oh my.  
You sure you want to?  
Yeah, it's OK if you want.  
That's try yes.  
If you want to give a brief update by voice, or do you have anything to show or?

  
Changjoon Park 1:06:58  
Yeah. OK.  
Yeah.  
OK then.  
Uh, first of all, I am.  
Yeah, responsible for API.  
The feature sets in trapping data sets and said I pointed out that about the 60,000 Support 6000 of data set from I speak a data from the data set and I applied applied machine learning mode as the in three machine learning mode as in tortas before I train the machine learning models, I separate the training set and they separate this data as the from the data training set and they test the set with 80% or 20% response respectively and I EI implemented food.

  
Fabio Pierazzi 1:07:10  
Like.

  
Changjoon Park 1:07:50  
I implemented the full machine learning model, which is like Decision tree as if we a linear as if we VM and naval values and last one it is CNN motors and I focused on the classifier provides the learning so these four are that my.  
The classifier model is so after training this data set, I evaluated with a some the evaluation functions like obviously curve and F1 score.  
I implemented the two these two and evaluated these result that training Wizard with two umm the the two depilation functions.

  
Fabio Pierazzi 1:08:35  
Like.

  
Changjoon Park 1:08:37  
So in conclusion, actually I'm the that it always occurs when I evaluate.

  
Fabio Pierazzi 1:08:42  
It's.

  
Changjoon Park 1:08:47  
Umm ourish.  
First of all, when I evaluate if Irish curve I evaluated with these move that with RC curve.  
In general, these models shows that one 0.9% of accuracies.  
Of course it's it depends on the models.  
For example, the decision trees on decision trees on average shows a true on the drill point.  
7% of accuracies where whereas the linear as.  
If we VM models show that shows that it drew point above.  
Peter appointed 9 nearly Abitur, appointed 9 accuracy of the accuracy for this training set and I tried.  
I implemented and the F1 score as well and it shows that it shows the different accuracy from the the curve and it evaluates its smooth as above.  
It's smooth as on umm.  
And it's like that drop .87 on average and I I was trying to keep on with these models with the shop they to explain.  
Umm yeah.  
To explains, I don't understand.  
I don't understand how to improve the design of detection algorithms, but and I yeah, I haven't.  
Yeah, it is.  
Diplomate did successfully not there.

  
Fabio Pierazzi 1:10:24  
OK.  
So, I mean in general I will show one, I think I mostly followed you, but when you have to give a verbal update, I think you tried to give me some numbers and colors and ideas.

  
Changjoon Park 1:10:34  
Yeah.

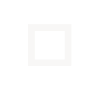
  
Fabio Pierazzi 1:10:37  
But I think saying it by voice is there to be too detailed.  
So sometimes when you give a verbal update, you may say even just try this classifiers and then I observed that the I don't know you're gonna say classifier say works better than B.  
Maybe there's a difference of 10%, but and then you were looking into shape for umm.  
For explaining also why the performance were that way, right?  
So as I mentioned to the others in this meeting, I guess the key point is beware of the base rate fallacy with the arc curves.  
So.  
In the autoship cars, you should really.  
So please read fallacies when there is imbalance that asset and especially in this case malware is the minority class.  
So if you do Arshin curve, you may get a very high performance, but it's since they're glasses are in balance that it's a little bit of an illusion.  
So you should really focus on the X axis on very small values of false positives, like between between 0 and 0.05 or 0 and 0.1.  
You may prefer to use procedure recalled curves.  
Beware also, and you say, I mean, I know that you will mention also score right when you.  
But you actually don't shock here, so I don't know if you're using accuracy as a metric.  
That's also very susceptible to resurrect fallacy.  
So you may get a very high accuracy of 90%, but on the performance on actually detecting malware maybe, but just very, very bad.  
So the F1 score where malware is the positive glass is usually better.  
So just just these NI guess, but in general it seems that you also started to do some experiments and you're investigating the results and you get a little bit of different performance.  
It's a little bit hard for me to give more detailed feedback because I not having the plots.  
I can't say much.  
Let's say right?  
But, but I think it's it's good progress.  
Keep in mind that for the preliminary report is going to be important also to.  
Maybe try should agree is going to be important as well, right?  
So it takes some time to be papers which self.  
It's expected that you also find some of your papers yourself.  
Some of papers or self, not just comment on mine and I think I shared with you also the how to review research for review blog post on my page right.  
I said something.  
I would review the system security measure.  
So yeah, we're gonna just share again here in the channel ohm, but keep in mind that for the preliminary project report, because it's going to be important until support positioning that your work as well.  
OK, I think we covered it.  
Apologies from my side because I also took some time in the beginning, maybe for next time.  
Let's try.  
I mean I I will try also to share the meeting better.  
Uh.  
From your side, try to keep.  
Right.  
Who?  
I won't Everage like 5 ish slides, right?  
So that you can stay within 7 minutes.  
I also didn't time very well, but let's try for next time to just stay within the hour also because usually in these meetings there are five students and we should stay within the hour.  
So we have well less, but I understand that she was partially my fault for sharing badly and starting off with this opening.  
But yeah, let's try just for next time to give it a more focus the and I know that some of you did a lot of slides.  
Just again try to keep it shorter.  
You don't have to give also basic definitions in these meetings.  
Like if you're presenting something based on a confusion matrix, you don't have to define it here necessarily, right?  
You may have to define something if it's very unconventional, and normally you read the this one paper and you came when you want to use that.  
But it's just that paper in the world that you set approach, you were very quick summary maybe.  
Well OK, it looks good in this group.  
So we have people that have started doing experiments and their numbers are already just great.  
So I mean, keep up with this and you're free, of course, to discuss with each other also for shop challenges.  
And if you don't have, I mean, if you don't have any other boarding question, I'll see you later today, so.

  
Jun Shi 1:15:42  
OK.  
Thank you, Sir.

  
Fabio Pierazzi 1:15:43  
Bye.  
See you.

  
Jun Shi 1:15:44  
Thank you.  
Bye.  
See you.

  
Yichen Zhou 1:15:46  
Hi.

  
Changjoon Park stopped transcription