**Interviewer 1:** [00:00:00] So yeah, if you just could confirm also on record that you agreed to the interview being taped.

**Interviewee:** [00:00:06] Uhum.

**Interviewer 1:** [00:00:09] So I'll start with some questions related to your background. Could you tell us what is your current position at your current job?

**Interviewee:** [00:00:17] Yes. I am a researcher at the research center called [Center Name: Removed for Anonymity] in [Country: Removed for Anonymity] that is, in which there are several universities that are involved and they work the for several [inaudible: stuff?], in particular my position regards to the use of big data and data mining methodologies for addressing different programs, for instance health or social network analysis. Typically we use different kind of deep learning application deep learning Technologies as you told before Keras, Pytorch or Tensorflow depending on the kind of application that we choose to use and depending on the particular function that we can find inside this framework.

**Interviewer 1:** [00:01:30] Thank you. So could you tell us what is your overall work experience? And what is your experience in terms of years in deep learning or machine learning?

**Interviewee:** [00:01:41] At least two or three years.

**Interviewer 1:** [00:01:44] Okay, thank you. And what type of deep learning networks have you developed, implemented and by that I mean like supervised or unsupervised, reinforcement learning?

**Interviewee:** [00:01:57] Okay, as a initially I worked about supervised learning and now my attention is focused on [probably told this by mistake as did not finished the word: unsupervised] reinforcement learning to improve the performance of our approach.

**Interviewer 1:** [00:02:14] Okay. And which problems were you trying to tackle using deep learning the network? An example could be like image classification speech recognition.

**Interviewee:** [00:02:24] I try to use the Deep learning Technologies for different instanses. First one concerns the use of deep learning for social network analysis in particular node classification. In other stuff you consider the use of deep learning for prognostic. Another one is a deep learning the use of the deep learning for image classification and the definition of decision support system. Another one is the use of deep learning for chat bot.

**Interviewer 1:** [00:03:05] Okay. You told us that you've been using Tensorflow, Keras, Pytorch and I'm guessing that as a language you have been using python so far or have used any other languages?

**Interviewee:** [00:03:18] Typically we use Python, but also we also investigate for Scala.

**Interviewer 1:** [00:03:25] Okay, so I think here general questions end. So, as I tried to explain in the general objectives of the interview, we have one general question for you, which is what type of bugs, problems, challenges have you faced while developing these systems. So if we could start from there and you could just list us whatever is on your mind regarding this question, that would be very useful.

**Interviewee:** [00:03:51] Typically the main programs that we have when we work with this kind of application concerns the lack of documentation in particular because it is mainly focusing on image analysis, an image classification. For this reason when we try to extend the evaluation for other stuff, for instance graph metric graph analysis, or short boat or whatever, It's difficult to adapt or to provide some metwork for this kind of application. Moreover, there are different stuffs, different implementation of a network for different framework. For instance, if you have an application for Keras it's really difficult to try to adopt this kind of application another framework.

For instance, if you use Keras and want to use Pytorch it is difficult to change this particular language, this particular network.

**Interviewer 1:** [00:05:06] Okay, Anything else?

**Interviewee:** [00:05:08] I don't think so.

**Interviewer 1:** [00:05:11] Okay. So when developing these systems have you used existing data sets to train your network or did you have to collect the data yourself?

**Interviewee:** [00:05:23] Both. Sorry one moment, please.

I try to use the some dataset that is possible to crawl from different websites because in this way it's easy to compare my results with other people, but also working on social network analysis. I have to crawl myself the information from the social networks.

**Interviewer 1:** [00:06:00] So could you tell us about the problems that you have faced while collecting the data yourself?

**Interviewee:** [00:06:06] The problem concerns typically the permission that I have to require to [from?] some people to crawl data, for instance if I use the Instagram or use Spotify or other stuff, I have to ask [inaudible] some people the possibility to gather all the information concerned.

**Interviewer 1:** [00:06:32] Is it like people at your University or?

**Interviewee:** [00:06:36] No, No, the idea is to start from them. And after that my idea is to ask [old, ??] people to provide some crawler or some request to them. It's also outside from the University.

**Interviewer 1:** [00:06:58] Yeah, could you tell us more about the technical problems that you face when you're collecting data?

**Interviewee:** [00:07:04] The problem concerns also the privacy problem, also the cleaning stuff. Because these data are usually require, often require to clean the data, to normalize the data and a [inaudible] identifying and providing way to store this data.

**Interviewer 1:** [00:07:28] Okay. So when using the existing datasets, did you face any problems when doing that?

**Interviewee:** [00:07:39] In this case, it's more easily to manage to handle because there are too much information that we can use to analyze all the information inside the data set. There are some normalization function to manage this data and typically are provided in some shareable format.

**Interviewer 1:** [00:08:05] Okay, so you mentioned that you have to clean data, etc. so these are all pre processing things. I wanted to ask if you ever had a case when you did not pre-process your data and then you faced problems when training and you had to go back and add some pre-processing.

**Interviewee:** [00:08:27] Do you talk about the possibility to not provide the any pre-process analysis?

**Interviewer 1:** [00:08:36] Not exactly. I'm just asking that if you ever had a case, when not doing pre-processing caused you problems for your training and you had to go back and add that pre-processing step.

**Interviewee:** [00:08:48] This is the case typically when I work with streaming process. In which it is possible to define some criteria to make.. one moment, please.

Okay, one moment. The problem concerns the the use of streaming data because when you define some pre-processing stuff, some pre-processing approach, typically, its required to update and to properly manage these stuff because data that you have is really heterogeneous and large in number. So it is required often the use the reinforcement learning approach to try to keep some useful information from this data.

**Interviewer 1:** [00:10:07] So it's not about pre-processing, about how to [ inaudible ] the reinforcement learning model.

**Interviewee:** [00:10:16] Uhum

**Interviewer 1:** [00:10:17] Ok, any other problems related to training data that you remember that you have, bugs, about technical problems, error messages.

**Interviewee:** [00:10:29] Okay, okay. Typically, it is due to social network analysis, is due the number of the requests that we try to ask for the server of the Twitter or other information that is related.

The fee that you have to pay for crawling this stuff. [inaudible] information, but there are also the possibility to obtain some timeout in your request due to the high latency on the network and also there are some particular problems related to the some special characters that you can find in some request..

**Interviewer 1:** [00:11:28] How do you handle special characters?

**Interviewee:** [00:11:33] Sorry?

**Interviewer 1:** [00:11:34] How do you handle special characters?

**Interviewee:** [00:11:36] Typically, depending on what these special characters are It is fine because you can find that inside the text for instance, a review or for instance a post, and in this case it's possible to use the Unicode format. And so you can manage properly this stuff, but it's also possible to use in some requests in which it's not possible to properly manage stuff.

So. For instance, the surname has some special character or the name, typical case - in the passport. And so in this case, we try to remove the suspicious character from the request and replace with some, you know, some loose character, some typical character.

**Interviewer 1:** [00:12:33] Okay. So you also mentioned that the training data can be very big, if I remember correctly. Tell us how you handle that.

**Interviewee:** [00:12:42] The amount of data?

**Interviewer 1:** [00:12:45] Yes.

**Interviewee:** [00:12:45] In terms of rule or in terms of size?

**Interviewer 1:** [00:12:51] Both.

**Interviewee:** [00:12:52] Okay. I think that in the first case we have as a size more or less 40 gigabytes. And has rule. It's really difficult to see because concerning this data set we use the music data set and we analyze the track. We manage them has images so more or less 20,000 images.

**Interviewer 2:** [00:13:29] Did you have any problems with managing this data and while training for example may be batching?

**Interviewee:** [00:13:39] Yes. This is typically is due to the hardware requirements required to properly manage this data because typically, when you manage this kind of data is large amount of data it is required to support the cluster of GPU's and so it's really difficult to find a cluster or it's really expensive. And there are moreover the some free tools who for instance Google Colab provide you some simple environmental that you can use but the problem is that only limited to 12 hours.

So the processing of the the data set is divided in different steps and you have to manage how do you update your network in different steps.

**Interviewer 2:** [00:14:43] And could you, please, tell us a bit in detail about the problems that you had, particularly?

**Interviewee:** [00:14:52] Yes, I think that for now the main problem concerns the hardware requirements, the lack of documentation and the difficulty to extend the existing network for other stuff and using in this case transferred learning. So it's really difficult in this case, but it's required to reduce your problem, to associate your problem with the well-known approach proposed in the literature.

And also because it's really difficult to try to define a new network so many people work directly on the existing one.

**Interviewer 1:** [00:15:48] So you told us that you collect some data from Spotify for example, and you collect music tracks. So what are you trying to achieve there? Are trying to classify this music track somehow or..?

**Interviewee:** [00:16:00] I try to classify them for recommendation of programs, but also to analyze user behavior.

**Interviewer 1:** [00:16:13] Okay.

**Interviewee:** [00:16:15] And for instance if we use Twitter, may be concern the sentiment analysis or opinion mining in streaming or image processing for video surveillance problem and so on.

**Interviewer 1:** [00:16:36] Okay. Maybe we could talk about the structure of the models that you've been using. So have you ever faced problems related to wrong model structure and by that we mean like the number of layers, type of layers, dimension layer and etc.

**Interviewee:** [00:16:53] Typically, we use the transfer learning so we only change the last layer of the network, but when we define a more specific approach, the problem concerned the number of the layer the width for the tuned layer the input layer and how to manage this way to inside the network because it's really difficult to try to define the width [?] of the hidden layer. But also to define the correct input format for your problem, the number of the layers, the parameters for each layer and also which kind of network it's better to use.

**Interviewer 1:** [00:17:47] Do you remember any specific problems that you had and you change something about the layer and your performance became better?

**Interviewee:** [00:17:58] For instance typically approach in deep Learning Works [inaudible] for classification problem, but we try to extend this network of a regression problem. And so the idea is to try to identify not the goodness or badness of a particular sample, but specific score for the sample that we analyzed.

**Interviewer 1:** [00:18:26] Could you tell us about the hyper parameter tuning did you ever had problems or bugs related to that?

**Interviewee:** [00:18:43] No, for the hyperparameters I haven't problem because many problem concerns the definition of the network and the choice of the input layer and the type of the layer [inaudible] network but for hyperparameters, I do not find any problem.

**Interviewer 2:** [00:19:08] But maybe you had some problems with batch sizes or batching?

**Interviewee:** [00:19:13] I think that it's been on the the kind of architecture that you have.

**Interviewer 1:** [00:19:19] Okay

**Interviewee:** [00:19:21] The problem concerns the type of the architecture the number of the GPUs and the size, dimension of your data set. So typically I change this parameter not depending on the particular problem, but based on given architecture.

**Interviewer 1:** [00:19:43] Does it apply also to things like learning rate number of epochs then?

**Interviewee:** [00:19:48] Yeah, and in this case I try to find a good way and analyzing the trend in the training phase. I try to define it as a given number of the epoch. Varying the learning rate to see how the network learn for the examined problem.

**Interviewer 1:** [00:20:12] But did You have cases when you had to change the value of this parameter - increase or decrease them based on your analysis, of course?

**Interviewee:** [00:20:20] Yeah. I use these analysis to find the best parameters.

**Interviewer 1:** [00:20:28] Ok, about a loss functions. Do you use predefined or a custom.. Did you ever use a custom written loss function?

**Interviewee:** [00:20:38] I try. Usually I try to use the some classical loss function and where it is not possible to I try to define some metrics, typically depending on the local sensitiveness that it's possible to define.

**Interviewer 1:** [00:20:56] So you had cases when the predefined loss functions were not working well for you, right?

**Interviewee:** [00:21:03] Yeah, for instance, typically depending on the particular method that I choose to use for instance, It's possible to not use a metric that works on the entire Matrix, but only on the specific side some specific element inside the Matrix for instance, when you use a sparse Matrix in this case, it's useful that you define some classical metrics that works on the inter metrics, but you only use specific metric that is useful for you.

**Interviewer 1:** [00:21:48] And you told us that you had some problems related to hardware. So, where do you train your models? And could you tell us about the bugs, errors, problems that you have encountered that are related to Hardware?

**Interviewee:** [00:22:04] Typically main problem concerns the use of the amount of ram that CPU has. Because typically the information that we analyzed is really big and is higher in number. So in this case typically we have to restart our analysis because the ram is full and we have some problem to manage this stuff.

**Interviewer 1:** [00:22:46] Do you take any steps before you start training to avoid this memory issues?

**Interviewee:** [00:22:59] The problem is the [inaudible] the the memory of the GPU problem because for instance we use typically Google Colab where there is a Tesla GPU that has 16 gigabyte of RAM and it's not enough for us, for our analysis. So we try to identify some some cluster of GPUs where it is possible to provide our analysis.

**Interviewer 2:** [00:23:45] Did you try using nodes with multiple gpus and splitting work?

**Interviewee:** [00:23:52] I haven't tried it because I don't have any cluster of the GPUs but my idea is to use them. In fact, I'm working on some distributed framework based on cluster of the GPU to hide the complexity of the programming issues.

**Interviewer 1:** [00:24:25] So could you tell us about how do you select features of the training data?

**Interviewee:** [00:24:34] Okay, typically, we start from the [inaudible] works. We analyze the problem we try to understand the main features that you can find the in similar papers. After that, we try to define it newer newer ones and try to see if this kind of features improve our accuracy. Also performing some hyper parameter optimization, but also to feature selection or feature extraction, also using some embedding technique to transform our features in different vector spaces.

**Interviewer 1:** [00:25:32] Did you ever have a case when you used the wrong set of features and then you had to go back and remove some of them or add new features? If yes, could give us like more specific about that.

Oh, yeah, we typically it concerns the use of features about fake news problem for instance, in which the use of static and dynamic features provide a little better results. The combination of this one, so we a different kind of the network in which there are some features that works well in respect to other one, but if you use a other network, you have to repeat your analysis because in this case it's better to use some other features in respect to the previous one.

**Interviewer 2:** [00:26:26] I just wanted to ask, did you ever face a case when you couldn't find enough information or you couldn't get enough information from the data set you had.

**Interviewee:** [00:26:42] Okay. Typically I found this problem in health application in which I have a subset of information and it is not enough to understand the problem, to obtain higher results, but this depend only on privacy problem - too difficult to retrieve and catch the useful information you want. It's not possible to find in different way other features because you have only the analysis that the patient made in this particular day. So you cannot ask to a [inaudbile] to provide better analysis [inaudible]. So in this case, it's really difficult to define main features.

**Interviewer 1:** [00:27:50] And in cases, when you had to collect your training data, did you have any cases when you have to manually label it or not?

**Interviewee:** [00:28:00] Yes. Typically it concerns the social network analysis because in this case we have really small dataset that have a label but when you increase the size of your dataset, it's really difficult to define the ground truth. So you can define a good standard that uses some well-known open source application for instance Amazon Turk or you can use also some combination of some parameters to define the relevance of given sample with respect to other ones

**Interviewer 1:** [00:28:43] In your experience with Amazon Mechanical Turk how often people mislabel things?

**Interviewee:** [00:28:51] Ooo, I think that it depend on the age of the people because if we try to use both Amazon Turk and some students inside our University, if we use some student inside our University and they haven't degree or maybe they have a BSc [not sure] degree, the error rate is more or less 20-50 percent because. Because you have to train them about the correct use and the correct label procedure. Instead, if you use Amazon Turk It's possible that you can find the many people that are more interested to obtain money with respect to provide a good classification and this case I think that error is quite similar to, its in the range of 10 to 50 percent.

Sorry, ten- fifty percent?

Sorry 10-15 percent.

**Interviewer 1:** [00:30:11] So I wanted to ask did you ever have a case when you use the training data as it is for the training then you thought that the results are not good or right and then you found out that you had some wrong data in your training data?

**Interviewee:** [00:30:32] Okay, typically it concerns I think [inaudible], but it do not concern the Amazon Turk because we have some samples, some features about each buttons, but we don't have any information about the particular class to this kind of people belongs so we use the information that came in from medical stuff, but it is still so possible that they provide some different evaluation information to classify them. So, using these different knowledge we reach the worst results because they are in contradictory among them. And so it's really difficult to use the their knowledge.

**Interviewer 1:** [00:31:47] Is there a way for you to deal with that? Can you ask people to correct the data or it is as it is and there's nothing more to do about it.

**Interviewee:** [00:31:58] The problem is concerned the asked to them to reclassify the data is a really big problem. And typically we try to ask to entire number of people in order to reduce this error rate.

**Interviewer 1:** [00:32:16] Okay. So in your experience is there like some error messages that appear too often for you.

**Interviewee:** [00:32:27] But in which, in which case?

**Interviewer 1:** [00:32:28] In any case, just in your general experience with Learning System.

**Interviewer 2:** [00:32:33] Any bugs that happened often?

**Interviewee:** [00:32:36] I think that main bugs concern the the use of Open Source software for instance software that researchers publishe on GitHub, for instance. There are basically is not provided any documentation about it. So if you try to use it, do you have you to understand any information inside to the software. So it's really difficult to try to use it directly from when you crawl it from the GitHub repository, for instance.

**Interviewer 1:** [00:33:13] For example, for me I get like tensor shape wrong censorship kind of messages quite often. So do you remember any this kind of bugs that happens?

**Interviewee:** [00:33:25] I don't remember this [inaudible] at this moment, I try to figure out but..

**Interviewer 2:** [00:33:31] Any bugs that you did yourself in your code?

**Interviewee:** [00:33:37] Oh too much bugs [everybody laugh] I think that it depend on maybe on the definition of algorithm. Typically, I have different bugs, for instance, how it is possible to manage the information especially related to the memory requirements because when you manage a large amount of data the one of the main problem is how do you store and how you put in memory all these data. And so typically the problem concerns thecorrect use of the data structure for manage this data.

**Interviewer 1:** [00:34:38] Anything else that you want to add?

**Interviewee:** [00:34:44] Sorry?

**Interviewer 1:** [00:34:45] Anything else you want to add?

**Interviewee:** [00:34:46] No, no. No, I don't think so.

**Interviewer 2:** [00:34:49] I have a little question, just to confirm. As I understood you had such cases when you had some problem to tackle you had the data and you chose on some network model like structure, whatever and then in the process you understood that this particular model doesn't work quite well for your task and you had to change it.

**Interviewee:** [00:35:12] I try different network because I think that it's better compare different approaches in this case. Typically I use the LSTM, RNN or classical artificial intelligence approach for research as SVM, Random Forest. I try to compare this kind of network and try to figure out which kind of network better fit on my program.

**Interviewer 2:** [00:35:42] Okay.

**Interviewer 1:** [00:35:45] So if you don't have anything to add I think we will let you go finally. Thanks a lot for your time.

**Interviewer 2:** [00:35:54] Thank you. And if you remember something by chance, or you get new bugs and you could just drop us a message about it. That would be cool.

**Interviewee:** [00:36:04] Okay

**Interviewer 1:** [00:36:06] Anytime you get a bug you should think about us and drop us a message. Okay. Thank you.

**Interviewee:** [00:36:17] Thank you, have a nice evening. Bye.