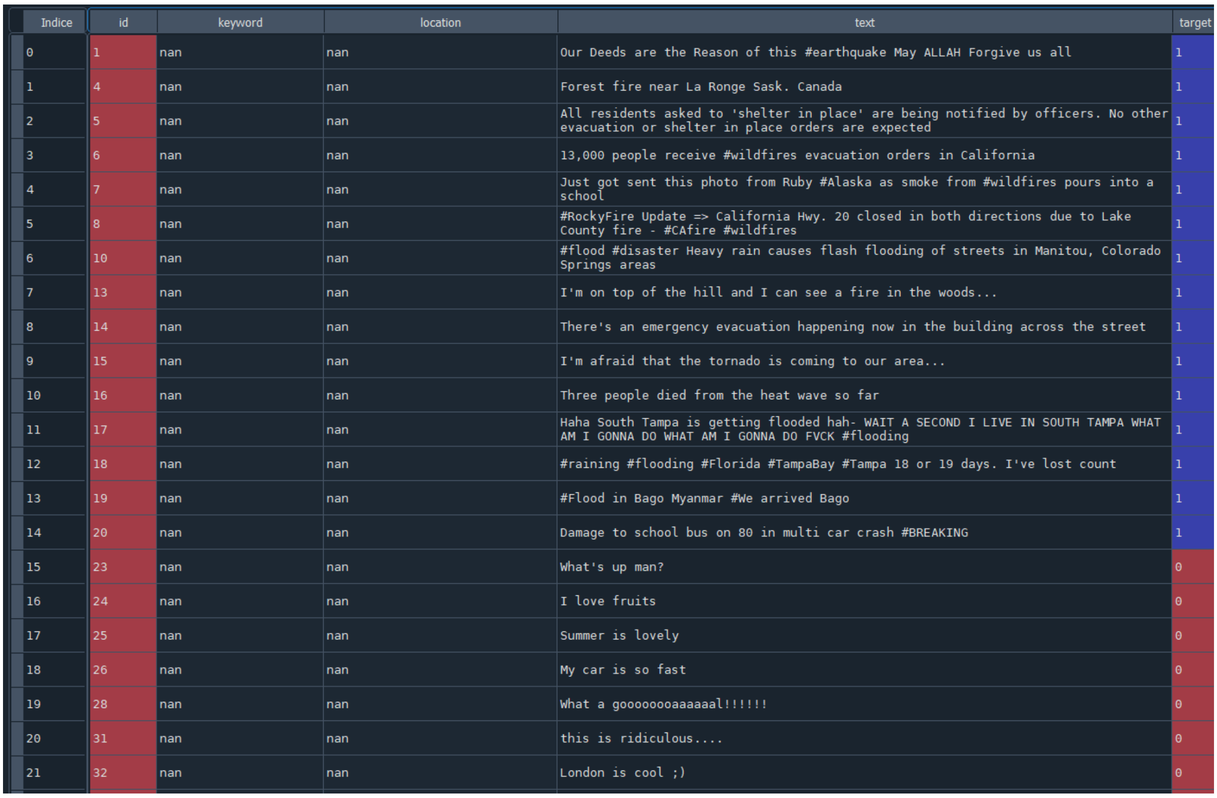
Munier Antoine

Homework 2

1. intro

The goal of this exercise is to compare two models (LSTM and GRU) for the classification of tweets that are or are not related to disasters. We will train both models on the "train.csv" dataset, which is taken from the Kaggle website. This dataset contains tweets classified into two categories: those that talk about disasters and those that do not.

representation of the dataset :



Indice	id	keyword	location	text	target
0	1	nan	nan	Our Deeds are the Reason of this #earthquake May ALLAH Forgive us all	1
1	4	nan	nan	Forest fire near La Ronge Sask. Canada	1
2	5	nan	nan	All residents asked to 'shelter in place' are being notified by officers. No other evacuation or shelter in place orders are expected	1
3	6	nan	nan	13,000 people receive #wildfires evacuation orders in California	1
4	7	nan	nan	Just got sent this photo from Ruby #Alaska as smoke from #wildfires pours into a school	1
5	8	nan	nan	#RockyFire Update => California Hwy. 20 closed in both directions due to Lake County fire - #CAfire #wildfires	1
6	10	nan	nan	#flood #disaster Heavy rain causes flash flooding of streets in Manitou, Colorado Springs areas	1
7	13	nan	nan	I'm on top of the hill and I can see a fire in the woods...	1
8	14	nan	nan	There's an emergency evacuation happening now in the building across the street	1
9	15	nan	nan	I'm afraid that the tornado is coming to our area...	1
10	16	nan	nan	Three people died from the heat wave so far	1
11	17	nan	nan	Haha South Tampa is getting flooded hah- WAIT A SECOND I LIVE IN SOUTH TAMPA WHAT AM I GONNA DO WHAT AM I GONNA DO FVCK #flooding	1
12	18	nan	nan	#raining #flooding #Florida #TampaBay #Tampa 18 or 19 days. I've lost count	1
13	19	nan	nan	#Flood in Bago Myanmar #We arrived Bago	1
14	20	nan	nan	Damage to school bus on 80 in multi car crash #BREAKING	1
15	23	nan	nan	What's up man?	0
16	24	nan	nan	I love fruits	0
17	25	nan	nan	Summer is lovely	0
18	26	nan	nan	My car is so fast	0
19	28	nan	nan	What a goooooooaaaaaal!!!!!!	0
20	31	nan	nan	this is ridiculous....	0
21	32	nan	nan	London is cool ;)	0

2. cleaning text

With a quick glance at the training dataset, we quickly notice that the data will need to be processed, as the tweets contain a lot of characters, formatting, and other elements that make it more difficult to extract the actual information. We therefore need to "clean" the tweets so that they all have a common structure, without the "artistic signatures" added by their authors.

Here are some examples:

- “ @bbcmtd Wholesale Markets ablaze http://t.co/lHYXEOHY6C ”
- “ 320 [IR] ICEMOON [AFTERSHOCK] | http://t.co/e14EPzhotH | @djicemoon | #Dubstep #TrapMusic #DnB #EDM #Dance #Ices□Û_ http://t.co/22a9D5DO6q”
- “ Traffic accident N CABRILLO HWY/MAGELLAN AV MIR (08/06/15 11:03:58)”

We quickly notice a few elements that can be removed. First, URLs, they are easy to detect, add nothing useful for classification, and likely confuse the model rather than help it understand the sentence. We also notice many hashtags (#); these can be removed, but the word following them should be kept. Indeed, hashtags often summarize the tweet, making them very useful. We can find hashtags like #wildfires or #earthquake, which directly help classify the tweet. Among other special characters, there's **@**, which is used to tag a person. Therefore, we don’t need to keep the word that follows an @ — we simply remove it. We also remove digits and numbers. The information helping us classify a disaster-related tweet might come *after* a number, but the number itself is never meaningful. Similarly, we remove special characters like %,&,=, +, and others.

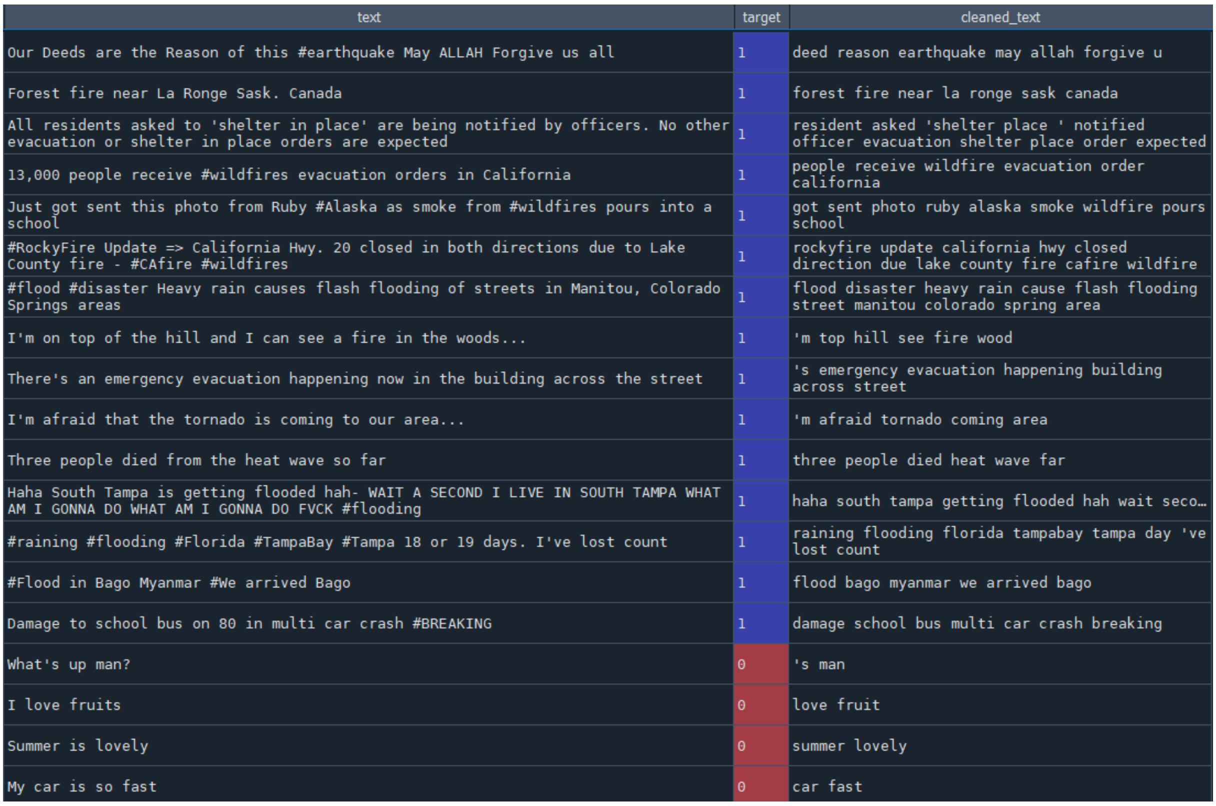
Due to the limited number of characters in a tweet, many people use abbreviations to convey their message. Therefore, our cleaning function must detect these abbreviations and replace them with the full word. This way, tweets containing abbreviations will be processed in the same way as those without. To do this, we use a list of abbreviations along with their meanings, and each time our function recognizes an abbreviation, it replaces it with the full version.

We also remove very common words. They make the sentences heavier to process without providing much information about whether the tweet is about a disaster or not. We keep only the less frequent words — and since disasters are not part of everyday life for most English speakers, the words related to disasters tend to be among those less common ones. As a reminder, the goal is not to understand the meaning of the sentence but simply to know whether it speaks of a catastrophe or not.

As a result, we get shorter sentences that still contain the key words needed for classification, which improves classification performance. We also replace multiple spaces with a single one, as this can slightly interfere with the learning process. We also remove punctuation, as it doesn’t help us determine whether a tweet is about a disaster or not.

In the Sample document provided to help us with this task, we found other functions, such as one to remove emojis. We chose not to use it because, after some research on the dataset, it seems there are none. So there is no need to make our function heavier. We will check if this is also the case in the test dataset. We have reduced our tweets to concise word lists that are easier to process and more likely to help with classification.

Results :



text	target	cleaned_text
Our Deeds are the Reason of this #earthquake May ALLAH Forgive us all	1	deed reason earthquake may allah forgive u
Forest fire near La Ronge Sask. Canada	1	forest fire near la ronge sask canada
All residents asked to 'shelter in place' are being notified by officers. No other evacuation or shelter in place orders are expected	1	resident asked 'shelter place ' notified officer evacuation shelter place order expected
13,000 people receive #wildfires evacuation orders in California	1	people receive wildfire evacuation order california
Just got sent this photo from Ruby #Alaska as smoke from #wildfires pours into a school	1	got sent photo ruby alaska smoke wildfire pours school
#RockyFire Update => California Hwy. 20 closed in both directions due to Lake County fire - #CAfire #wildfires	1	rockyfire update california hwy closed direction due lake county fire cafire wildfire
#flood #disaster Heavy rain causes flash flooding of streets in Manitou, Colorado Springs areas	1	flood disaster heavy rain cause flash flooding street manitou colorado spring area
I'm on top of the hill and I can see a fire in the woods...	1	'm top hill see fire wood
There's an emergency evacuation happening now in the building across the street	1	's emergency evacuation happening building across street
I'm afraid that the tornado is coming to our area...	1	'm afraid tornado coming area
Three people died from the heat wave so far	1	three people died heat wave far
Haha South Tampa is getting flooded hah- WAIT A SECOND I LIVE IN SOUTH TAMPA WHAT AM I GONNA DO WHAT AM I GONNA DO FVCK #flooding	1	haha south tampa getting flooded hah wait seco…
#raining #flooding #Florida #TampaBay #Tampa 18 or 19 days. I've lost count	1	raining flooding florida tampabay tampa day 've lost count
#Flood in Bago Myanmar #We arrived Bago	1	flood bago myanmar we arrived bago
Damage to school bus on 80 in multi car crash #BREAKING	1	damage school bus multi car crash breaking
What's up man?	0	's man
I love fruits	0	love fruit
Summer is lovely	0	summer lovely
My car is so fast	0	car fast

* Our function does not replace the original tweet with its "corrected" version but adds a new column

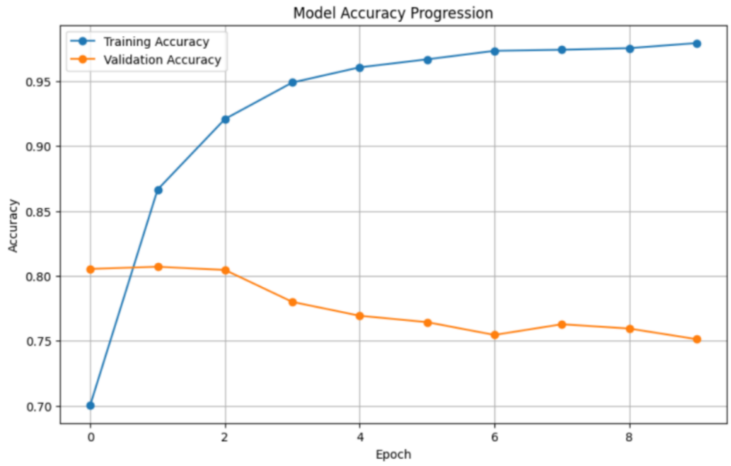
3. Models

3a. Split and tockenisation

To evaluate our two models we will train them and evaluate them on the "train" dataset. We therefore separate it into 2. We tokenize the dataset with the Keras tokenizer and once this is done the text is ready to pass through the LSTM and the GRU.

3b. LSTM

For the LSTM model, we simply take the LSTM from the sample document. Without changing the hyperparameters, here is the result.

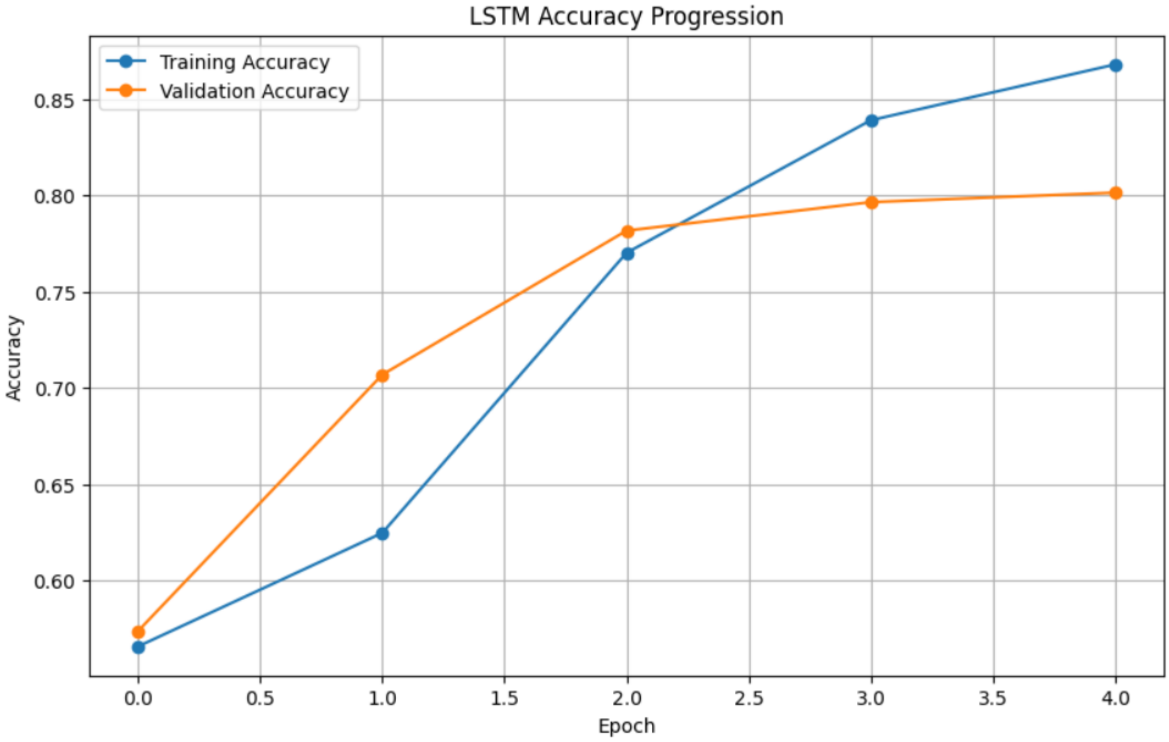


We note that the model converges quickly (5 epochs) towards its final accuracy on the training data. However, we note a (decreasing) stability on the evaluation data. This indicates a difficulty in generalizing the model. The difference between the two curves is a sign of overfitting (in addition to the decrease in validation accuracy).

To correct this, we can already reduce the number of epochs. Indeed, from 5 epochs, our model converged and only overfits afterward. We will also try to adjust the hyperparameters to see if we can improve the results.

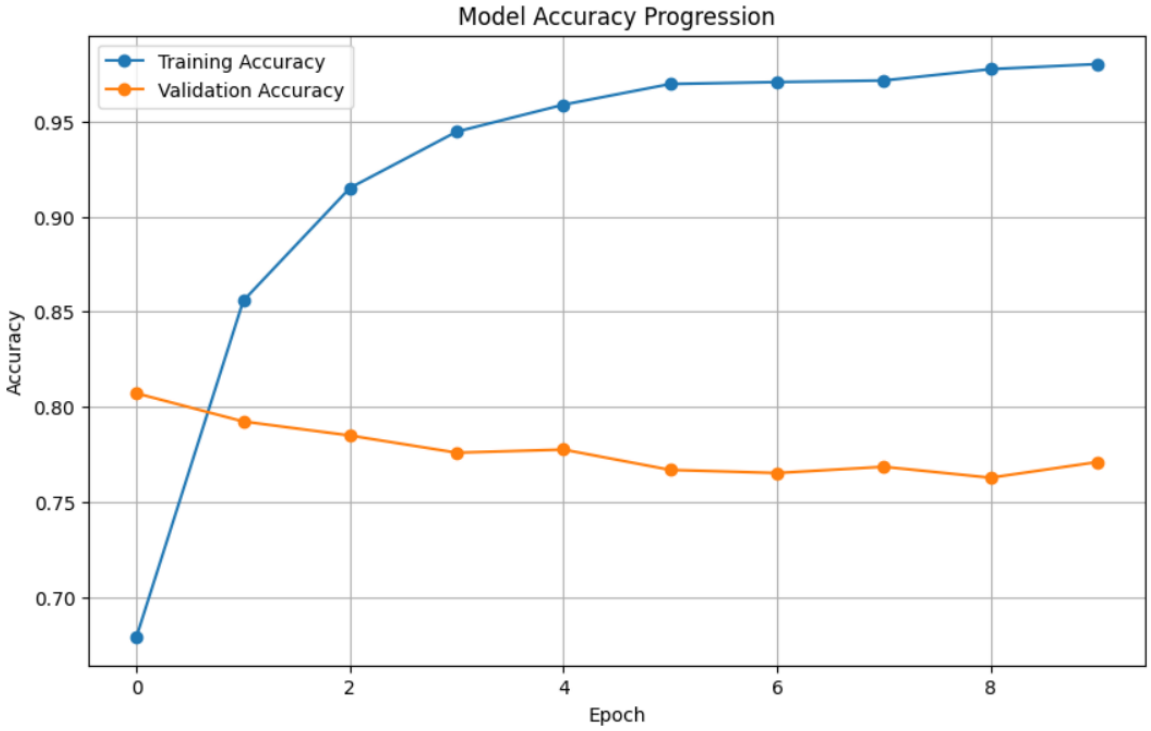
Reducing the batch size makes the model converge more quickly toward its final accuracy. This is normal because the gradient update depends on fewer examples.

After modifying the hyperparameters, here is the new curve:



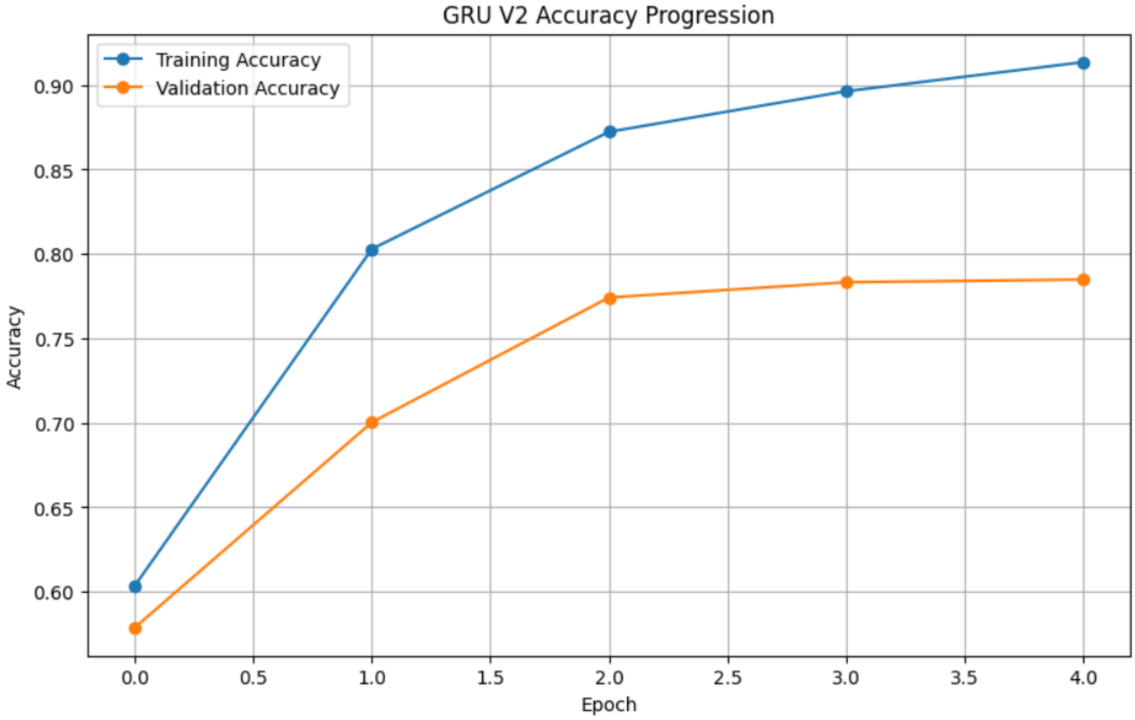
3c. GRU

To build our GRU, we first started with a simple version to see the results. Here they are:



We notice that this model also seems to overlearn. It converges in 5 epochs on the training data and is stable (in decline) on the evaluation dataset. We tried to play with the hyperparameters, but we didn't necessarily get better results (validation accuracy curve growing and closer to the training curve). We then moved on to a more complex model by adding a bidirectional to our model. This will unfortunately increase the calculation time but in this phase our only criterion is accuracy (and not overfitting). We also use the L2 function to penalize excessively large weights, this further reduces overfitting. We also increase the dropout and recurrent_dropout to further avoid excessively large weights and strong dependencies of the model at certain connections. In the same logic, we use Batchnormalization which stabilizes the training and makes it faster.

Results :



The learning curve on both is more "pretty". We see progression for both, there is still a gap between them but it is a little smaller. We also see that we did well to stop at 5 epochs because we are already stagnating. (In the final test we will train on the entire train dataset, so there will be more data for training). The model is therefore better, not necessarily in accuracy but not overfitting. It is longer than the Anscien but it is less important.

4. Comparison of models

To compare our models, we chose the following six criteria:

Accuracy: The main criterion for the classification accuracy of our two models.

Loss: A measure of the error during model training. This allows us to observe the model's stability.

Training time: It's important to know how long it takes for our model to run.

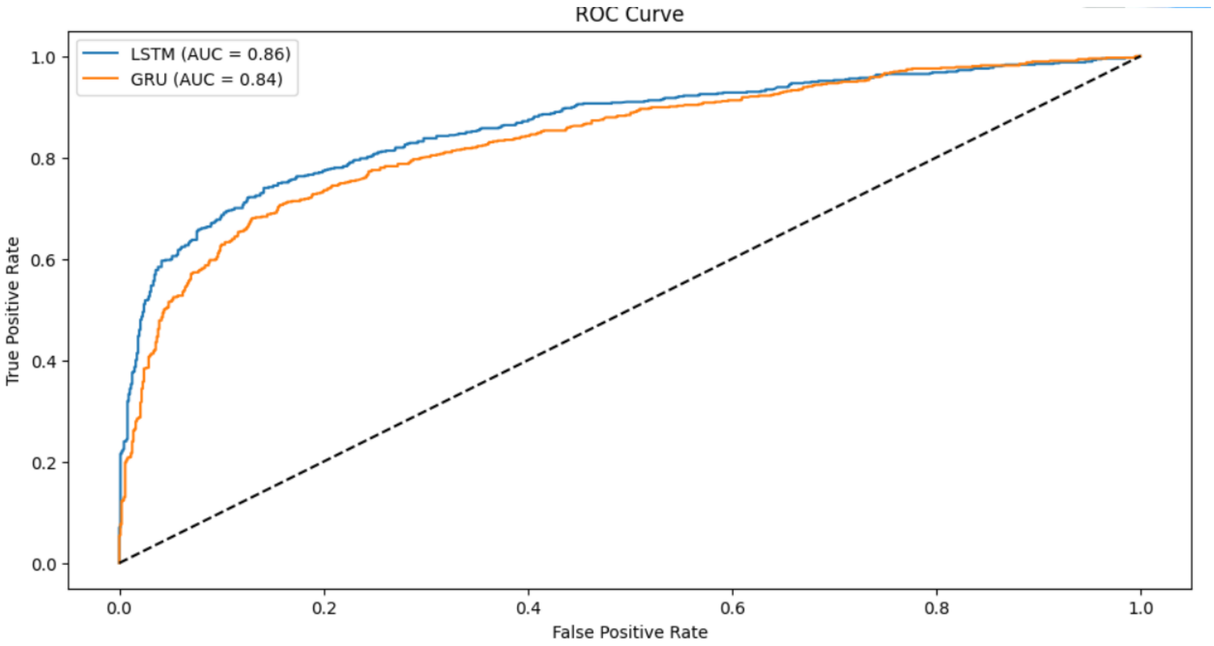
Inference time: We also measure how long it takes for the model to answer a question.

Model complexity (number of parameters): It's important to know how complex our model is.

ROC curve: When testing classification models, it's common to plot the ROC curve. It's directly related to accuracy. (I also include it to provide another curve.)

Résults :

Critera	LSTM	GRU
Accuracy	0.805	0.771
Loss (écart-type)	0.023	0.079
Training time	37.5s	48.46s
Inference time	0.58s	1.47s
Nulber of prameters	80553	327313



The LSTM is better on all criteria; there seems to be no real debate.

7. Testing

For the testing part, we will retrain the models on the entire train dataset, then we will generate the response csv on the test dataset in the requested format.