# Automatically update ML model on tweet sentiment analysis on the coronavirus crisis

This project will be an INDIVIDUAL research project.

I am a non-local student from France. I am living in France and the country is currently on lock-down. Because of this I will not be able to come back to Hong-Kong before May. In case of the university re-open I didn’t want to be part of a group and jeopardize the work of other by being absent. This is why I decided to work individually.

The objective is to leverage the cloud native service to improve our own ML model. AWS Comprehend is a fully managed NLP service, we can use it to understand and extract sentiment analysis. However this service is not fully tailored to your own need. In this project we will use this service and our own model. The objective is to combine both in order to improve our own model. By using both we will try to create a new dataset automatically which will grow with time. Our own model will retrain automatically overtime in order to improve it’s accuracy. A company could possibly use this system in order to improve its own model in the beginning. Then once it’s dataset is big enough and it’s model deemed good enough it could stop using the cloud service and keep using its own ML model.

We will create an ML model with insight on the topic that we will analyze here the coronavirus crisis.

We will create an automated script that will aggregate the result to create our new Labelled Dataset (composed of the tweets that will have obtained the same label in our model and in AWS Comprehend).

We will create a script to automatically train our ML model again.

The project will follow this architecture:

AWS Lambda (Server Less Service) will be used to query tweets (twice a day for example). Then AWS Kinetic will be used to preprocess tweets. Those data will be send to AWS Comprehend to extract information and will also be stored in AWS S3. On an AWS EC2 instance we will run our ML model. Over time once we will have enough result Compare both result. When the result are the same label the tweet. Create new Dataset. Train automatically our model again.