Introduction

The IBmer WatsonX Challenge was an amazing opportunity to improve our skills in this fantastic framework. Our team, the Brazilian AI IBM Mavericks, choose the track 2 - Determine customer satisfaction with watsonx.ai to start dealing with foundation models and develop concepts to better understand customer satisfaction.

Preparation

With the objective of share knowledge with all team members our 1st strategy was that each one was supposed to explore the WatsonX Challenge Sandbox and deep dive in this arriving technology.

Strategy

As a planned, it was agreed that each participant in the group should experiment and create their own prompts and try to get good results on the classification tasks and then use the best results to calculate the scores.

Execution

Here we have to distinct results, the 1st one got from the original prompt used to identify sentiment of customer opinion about Service.

```
[42]: print(satisfaction_instruction)

    Find the sentiment of the Customer in the text. Generate 0 for not satisfied, 1 for satisfied

    comment: I have had a few recent rentals that have taken a very very long time, with no offer of apology.  In the most recent case, the agent subsequently offered me a car type on an upgrade coupon and then told me it was no longer available because it had just be

    satisfaction: 0
```

In this case in the first we got an F1 of 0.96, and we concluded that we had a good classification model.

Calculate the F1 micro score

```
[16]: from sklearn.metrics import f1_score

      print('f1_micro_score', f1_score(satisfaction, results, average='micro'))

      f1_micro_score 0.96
```

The second activity was to classify the business area related to such opinion. Which ones were related to Products or Service and their sub classifications.

After the team had experimented several configurations and several prompts, we noticed that the results weren't good enough, obtaining a F1 score next to 0.06.

```
[59]: print(business_area_instruction)
```

```
   Find the business area of the customer e-mail. Choose business area from the following list: 'Product: Availability/Variety/Size', 'Product: Functioning', 'Product: Information', 'Product: Pricing and Billing', 'Service: Accessibility', 'Service: Attitude', 'Service: Knowledge', 'Service: Orders/Contracts'.


   comment: I have had a few recent rentals that have taken a very very long time, with no offer of apology.  In the most recent case, the agent subsequently offered me a car type on an upgrade coupon and then told me it was no longer available because it had just be

   business area: 'Product: Availability/Variety/Size'
```

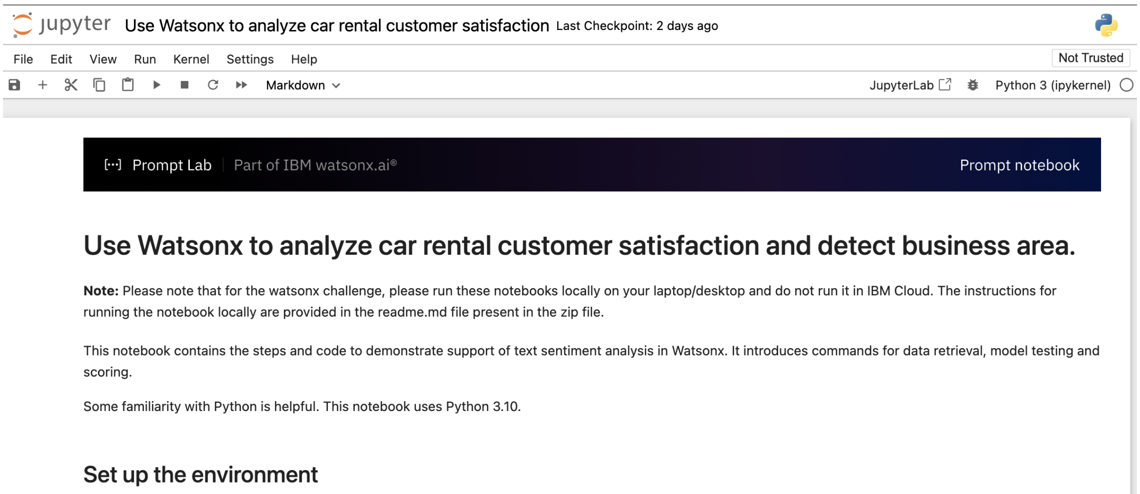
Calculate the F1 micro score

```
[65]: from sklearn.metrics import f1_score

      print('f1_micro_score', f1_score(area, results, average='micro'))
```

```
f1_micro_score 0.06
```

Then we changed the strategy and started to use the jupyter notebook provided to simulate combinations of examples and try to find the best one to evaluate the model.



As suggested in the notebook instructions, we created several samples as can be seen here:

2. Business area

Define instructions for the model to detect business area in which the customer review is targeted.

Note: Please **start with using** watsonx.ai Prompt Lab to find better prompts that provides you the best result on a small subset training records (under `train_data` variable). Make sure to not run an inference of all of `train_data`, as it'll take a long time to get the results. To get a sample from `train_data`, you can use e.g. `train_data.head(n=10)` to get first 10 records, or `train_data.sample(n=10)` to get random 10 records. Only once you have identified the best performing prompt, update this notebook to use the prompt and compute the metrics on the test data.

Action: Please edit the below cell and add your own prompt here. In the below prompt, we have the instruction (first sentence) and one example included in the prompt. If you want to change the prompt or add your own examples or more examples, please change the below prompt accordingly.

[73]:
```
# all
#sample_train = train_data

# 1st 10
#sample_train = train_data.head(n=10)

# random 10
#sample_train = train_data.sample(n=10)

# random 10 prop
#sample_train = train_data.groupby('Business_Area').apply(pd.DataFrame.sample, frac=.1)

# best
sample_train = train_data.filter(items=[73, 60, 31, 86, 43, 21, 56, 47, 98, 45], axis=0)

sample_train[["Customer_Service","Business_Area"]]
```

[73]:

	Customer_Service	Business_Area
73	If there was not a desired vehicle available t...	Product: Availability/Variety/Size
60	Very good. willing to drop me off and pick me ...	Product: Functioning
31	delayed shuttle, almost missed flight, bad cus...	Product: Functioning
86	I thought that they were very short and not ve...	Product: Functioning
43	My last experience was seamless. The only th...	Product: Pricing and Billing
21	they should not try so hard to up sell	Service: Attitude
56	Our customer service representative was polite...	Service: Knowledge
47	it was fine.	Service: Knowledge

This way, we improved the result achieving a F1 score of 0.3 what, once again, wasn't a good result to prediction task.

[83]:
```
business_area_instruction =\
  "Find business area on comment. Choose business area from the following list: 'Product: Availability/Variety/Size', 'Product: Functioning'
  "comment: I have had a few recent rentals that have taken a very very long time, with no offer of apology. In the most recent case, the ag
   examples  #+ examples_know + examples_func

print(business_area_instruction)
```

```
Find business area on comment. Choose business area from the following list: 'Product: Availability/Variety/Size', 'Product: Functioning',
'Product: Information', 'Product: Pricing and Billing', 'Service: Accessibility', 'Service: Attitude', 'Service: Knowledge', 'Service: Orde
rs/Contracts'.

comment: I have had a few recent rentals that have taken a very very long time, with no offer of apology. In the most recent case, the agen
t subsequently offered me a car type on an upgrade coupon and then told me it was no longer available because it had just be
business area: 'Product: Availability/Variety/Size'

comment: If there was not a desired vehicle available the reps explored all options including competitors to assist in finding an available
vehicle. This level of service brought me back not to their competitor but the company as this reflects on their overall quality.
business area: 'Product: Availability/Variety/Size'

comment: Very good. willing to drop me off and pick me up from location, upgraded me to higher level due to availability
business area: 'Product: Functioning'

comment: delayed shuttle, almost missed flight, bad customer service
business area: 'Product: Functioning'

comment: I thought that they were very short and not very friendly. I felt like they hated their job and could care less about the custome
r.
business area: 'Product: Functioning'

comment: My last experience was  seamless.  The only thing I didn't like was having to fill the tank with gas before turning it in.  It was
inconvenient, but I didn't want to pay the hefty fill-up charge to the rental company.
business area: 'Product: Pricing and Billing'

comment: they should not try so hard to up sell
business area: 'Service: Attitude'

comment: Our customer service representative was polite and well-dressed. He smiled appropriately and answered my questions, not from a reh
earsed script, but from his own frame of reference. He shirt was neatly pressed and his hair was professionally coifed.
business area: 'Service: Knowledge'

comment: it was fine.
business area: 'Service: Knowledge'

comment:  The service was polite and professional. I was attended to quickly and courteously.
business area: 'Service: Knowledge'

comment: adequate to the price. It was fine
business area: 'Service: Knowledge'
```

Calculate the F1 micro score

```
[88]: from sklearn.metrics import f1_score

print('f1_micro_score', f1_score(area, results, average='micro'))
f1_micro_score 0.3
```

With these results we realized that maybe there was something wrong with the data sources, then we did some deep analyses to discover what was going on and the cause of so poor results.

Our conclusion was that the data sources were annotated mismatching the classifications which was leading to the poor results we got.

To prove this hypothesis, we modified the original data sources provided by organization, re-annotating some classifiers, that we noticed were more deviated: Product: Functioning and Service: Knowledge.

With these new data sources we achieved a F1 result of 0.4 what takes us to really assume that there are a problem with the business areas classification.

Calculate the F1 micro score

```
[97]: from sklearn.metrics import f1_score

print('f1_micro_score', f1_score(area, results, average='micro'))
f1_micro_score 0.4000000000000001
```

Conclusion

We liked very much of this initiative, but more than this, we found that the IBM strategy to work with trustable data is very important to successful projects.