PHARMA HACKS 2022

CHALLENGES

SSMU MLH McGill Department of Pharmacology and Therapeutics Pfizer

medtech TALENT ACCELERATOR IRCM INSTITUT DE RECHERCHES CLINIQUES DE MONTRÉAL Phyla

EVENT SCHEDULE

SATURDAY, MARCH 26

10 AM - Opening Ceremony
11 AM - Challenges Released, Hacking starts
11 AM -12PM - Venue opens for check-in
1-2 PM - MedTech Info Session
2-3 PM - Workshop: Data Management in Python
3-4 PM - Workshop: Introduction to Machine Learning in Python
8 PM - MLH Activity: MS Paint Bob Ross painting exercise
8:30 PM - All attendees leave venue
9 PM - Venue closes

SUNDAY, MARCH 27

10 -11 AM - Venue opens
11 AM - Challenges due
4:30 PM - Closing Ceremony
5 PM - All attendees leave venue, venue closes

PHYLA CHALLENGE ONE

INFORMATION EXTRACTION

Background

The amount of publications keeps growing steadily in the scientific literature, making it increasingly difficult to keep up with the knowledge in the biomedical and health fields. The quantity of freely available information combined with the lack of automatic archiving of obtained results introduces the need to automatically structure textual biomedical data and extract results for further usage. Natural Language Processing (NLP) methods provide a means for performing information extraction from such large sets of data publications.

Information Extraction

The process of extracting relevant information from a large number of unstructured texts is referred to as Information Extraction (IE). This process usually comprises two main steps: Named Entity Recognition (NER) and Relation Extraction (RE).

NER is the task consisting in identifying and categorizing specific entities in text. Examples of such entities include: disease, food, dietary supplements, symptoms, etc. RE is usually preceded by NER and consists in classifying the type of relation for a pair of entities found within the same sentence/paragraph. RE is often used for populating relational databases and knowledge graphs, which are useful to present information to users via search engines.

Task

Imagine you are a data scientist hired by a nutritionist who is trying to stay up to date on medical literature and who wants to know, for example, which foods can help or worsen certain symptoms of digestive illnesses like Irritable Bowel Syndrome (IBS) or Crohn's Disease. This information is extremely relevant for the nutritionist to be able to recommend a personalized diet to their patients. However, it is simply impossible for the nutritionist to keep track of relationships between foods and symptoms with the rate at which scientific literature grows. This is where you come in!

As a data scientist, you already have conducted the NER task and were able to retrieve sentences with a corresponding pair of entities. That's great but you have only done half of the work. In order to provide real value to the nutritionist, you now want to be able to detect the type of relation between the pair of entities in each sentence.

How?

You will have to build and train a machine learning model (of your choice) to classify each sentence and its corresponding entities pair in one of the three following categories (our target variable):

1. **Positive:** entity A has a positive impact on entity B
- Example: "Peppermint oil has been observed to reduce abdominal pain in IBS patients"
2. **Negative:** entity A has a negative impact on entity B
- Example: "FODMAP diets are highly associated with an increase in bloating among the patients."
3. **Not related:** entity A has no direct impact on entity B (or vice-versa)
- Example: "Bloating and abdominal pain are some of the common symptoms observed among IBS patients."

Dataset (here)

You will be provided with a dataset to work with. Each example will consist of the raw sentence, the pair of entities found in that sentence, the entities spans (their location in the sentence) as well as the corresponding target variable value (positive, negative or not related).

Model

You may choose the machine learning model you find best suited for this task. Keep in mind that some models might be more computationally intensive than others (language models such as BERT, or pre-trained versions such as Bio-BERT for example). As a data scientist, it is your responsibility to take this into consideration.

Training

When training a machine learning model, we often split our data in a train, validation and test sets. As a data scientist, it is your responsibility to design your experiments in order to achieve the best possible performance on your test set. Remember that the goal in machine learning is to develop a model that generalizes well on previously unseen data.

Result

Inference

You will then use your trained model to perform inference on an external test set (this is different from the test set YOU will define as part of the training), which will be provided to you without the corresponding relation types. Your model's performance will be evaluated based on the external test set. Remember that we are seeking good generalization!

Metrics of Scoring

The performance of your model will be evaluated on three different metrics

1. Precision
2. Recall
3. F1-score

More emphasis will be put on the f1-score. Can you see why?

Deliverables

You will be asked to submit:

1. Your predictions on the test set.
2. The model used to make predictions on the test set.
3. The notebooks, scripts used for the task (along with some explanation on how to run the code)
4. A report (between 1 and 4 pages)
    a. Exploratory data analysis/Data preparation
    b. Model(s) chosen and why? Don't describe the math behind the models, but simply explain why you tried this/these models.
    c. Training process: train/validation loss, confusion matrix, etc. (it's really up to you!)
    d. Train/test results using the metrics described above.
    e. A short discussion of the results.

Tips!

1. Make sure your model does not overfit your data. You might want to plot your training and validation losses!
2. Check whether the data is imbalanced or not. This could have an impact on the performance of your model. If it is, how can you deal with it?
3. Perform an error analysis. Is your model performing similarly for all classes (positive, negative, not related)?

Scoring Rubric:

Category	Description	Score
Documentation	• Did the team describe their exploratory analysis? • Did the team indicate why they chose a specific model? • Did the team describe how they evaluated the performance of the model (validation curves, confusion matrix, etc)?	1 2 3 4 5
Performance	• How good were the results compared to other teams?	1 2 3 4 5
Code	• Is the code documented? • Is the code "clean"? • Is the code organized?	1 2 3 4 5
Presentation	• Did the team clearly explain how they solved the challenge?	1 2 3 4 5

PHYLA CHALLENGE TWO

MULTI-DISEASE CLASSIFICATION

Background

Within the human gut there are trillions of microorganisms, referred collectively to as the gut microbiome, that contribute to many aspects of human health. The gut microbiome influences a wide range of human health, from food digestion to function of the immune system and mental health. The large influence the gut microbiome has on human health means that shifts in the microorganisms present can actually alter human health and lead to the development of some diseases. One goal in the medical community is to be able to diagnose different diseases based on the type and amount of bacteria present in a patient's gut microbiome. Several research groups have demonstrated the ability to accurately classify diseases, such as inflammatory bowel disease, depression, colorectal cancer, and Parkinson's disease.

The Problem

Commonly, researchers assess disease classification for one disease at a time. The training dataset contains gut microbiome information from patients with the disease and a similar number of healthy individuals. This leads to the potential that the machine learning model is simply recognizing the gut microbiome as "unhealthy" instead of identifying a specific signature for each disease. For example, a machine learning model trained to classify inflammatory bowel disease would also classify a patient that has colorectal cancer as having inflammatory bowel disease. The misclassification substantially reduces the ability to use machine learning models to classify different diseases in a clinical setting.

The Project

Develop a multi-label classification model to classify different diseases based on the microorganisms in the gut microbiome.

Dataset (here)

The dataset consists of the microorganisms in the gut microbiomes of people with different diseases. There are 1949 samples for Disease-0, 1213 samples for Disease-1, 578 samples for Disease-2, and 3741 healthy samples. Each row corresponds to a sample collected from an individual and each column refers to a bacteria identified in at least one person's gut microbiome. The "disease" column indicates the label for each sample.

Tip! The sum of all the microorganisms in each sample can vary substantially. This variation is not biologically significant and in fact an artifact of the technology used to collect the data. You will need to implement a way to transform the data of each sample to remove this artifact.

Metrics of Scoring

For the performance of your models to be assessed, report the following metrics:

1. F1 score
2. Cohen's kappa

Scoring Rubric:

Category	Description	Score
Innovation	• How unique or creative is the model design implemented? • How did the group account for the unbalanced dataset?	1 2 3 4 5
Documentation	• How well documented is the submitted code? • Is the code easy to follow?	1 2 3 4 5
Performance	• How good was the disease classification for the final model generated?	1 2 3 4 5
Code	• How organized is the submitted code? • How concise is the code for building the model?	1 2 3 4 5

MEDTECH CHALLENGE

MEDTECH TALENT ACCELERATOR CHALLENGE:

Background

Predicting how payments to physicians and teaching hospitals improve the probability of business success in the pharmaceutical and medical technology industries

Goal

- Build a database of pharmaceutical and medical technology companies that combines market capitalization and payments to physicians and teaching hospitals.
- Estimate the effect of payments as a leading or lagging indicator on % change in market capitalization

Data

- Open Payments https://openpaymentsdata.cms.gov/ (search by Company)
- Yahoo Finance https://ca.finance.yahoo.com/industries/healthcare/ (or similar healthcare finance database)

Scoring Rubric:

a) Presentation Skills

Criteria	Description	Score
Verbal Component of the Pitch	• Speakers are fluent and knowledgeable • Speakers are able clearly articulate concept • Appropriate use of vocabulary at a general level	1 2 3 4 5
Component of Delivery	• Enthusiasm and ability to convince • The audience feels engaged and the presenter is comfortable on stage • Appropriate use of gestures and body language	1 2 3 4 5
Use of Visuals and Text	• Good use of text • Good use of visuals • Organized, and easy to follow with minimum technical errors	1 2 3 4 5

b) Strength of the Proposed Idea

Criteria	Score
Potential for Technological/Knowledge Impact to Canada	1 2 3 4 5
Originality and Innovation	1 2 3 4 5
Extent to which the Scope of the Proposal Address All Relevant Issues	1 2 3 4 5
Clarity and Appropriateness of Methodology	1 2 3 4 5

PHARMAHACKS CHALLENGE

Goal

Use computer science to solve a problem in the pharmaceutical, medical, and biotech industries.

Scoring Rubric:

Criteria	Points
Creativity & Innovation	**10**
Real-life application	**10**
Clarity of presentation	**10**
Technical complexity	**5**

DISCLAIMERS

Thank you for your participation!

By participating to this hackathon, you understand and agree that:

(a) our challenges are confidential and must not be copied, shared or disseminated in any manner whatsoever;

(b) any information or work you will submit is yours e.g. does not violate any other person's IP rights and

(c) you will retain any of your IP rights and allow us as a challenge provider, without any further compensation, to use all work derived or relating to our challenges for our own purposes