Covid-19 Open Research Dataset Challenge (CORD-19)

Using NLP to Answer High-Priority Scientific Questions Related to COVID-19

Joseph Cheng

June 13th, 2020

Table of Content

Proposal

Domain Background

This project is derived from the Kaggle challenge[1] initiated by the White House, AI2, CZI, MSR, Georgetown and NIH.

The goal of the Call to Action by the White House[2] is to collaborate between the artificial intelligence community and the medical research group to come up with new text and data mining techniques which can help answer high-priority scientific questions related to COVID-19.

The main motivation of the project is to demonstrate how machine learning can accelerate the understanding of COVID-19 so that we can reduce, prevent, and ultimately remove the impact of COVID related pandemic to human health.

Problem Statement

With the rapid increase in Coronavirus literature, it has become increasingly difficult for the research community to keep up with the latest information. There are many high-priority scientific questions about COVID-19 that are awaiting to be answered. The task within the scope of the project is to create summary tables that address relevant factors related to COVID-19.

More specifically, the goal is to categorise each of the medical literature into the following research topics:

- Effectiveness of case isolation/isolation of exposed individuals (i.e. quarantine)
- Effectiveness of community contact reduction
- Effectiveness of inter/inner travel restriction
- Effectiveness of school distancing
- Effectiveness of workplace distancing
- Effectiveness of a multifactorial strategy prevent secondary transmission
- Seasonality of transmission
- How does temperature and humidity affect the transmission of 2019-nCoV?
- Significant changes in transmissibility in changing seasons?
- Effectiveness of personal protective equipment (PPE)

[1] "COVID-19 Open Research Dataset ...." 10 Jun. 2020, https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge. Accessed 15 Jun. 2020.
[2] "Call to Action to the Tech Community on New Machine ...." 16 Mar. 2020, https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/. Accessed 15 Jun. 2020.

By categorising research literature into the high priority scientific questions, it would accelerate researchers ability to perform hypothesis, experimentation, validation, and findings made by experts around the world.

Datasets and Inputs

The dataset used in this project is provided by The Allen Institute for AI, Chan Zuckerberg Initiative (CZI), Georgetown University's Center for Security and Emerging Technology (CSET), Microsoft, and the National Library of Medicine (NLM) at the National Institutes of Health. The leading research groups have prepared the **COVID-19 Open Research Dataset (CORD-19)** with over 138,000 scholarly articles about COVID-19, SARS-CoV-2, and related coronaviruses.

Data Dictionary

The source dataset contains the following files

Document Parse

- CORD-19 contains machine-readable Coronavirus literature ready for data mining techniques, which will be used for training and testing Natural Language Processing models in this project.
- All files is stored as JSON format under the folder pdf_json and pmc_json
- Each medical literature contains paper_id, metadata, references, citations, as well as contextual information such as title, text, and abstract.

Metadata

- The metadata contains basic information about each medical literature stored in CORD-19, including the references and citations.
- The Metadata will be a directory used within this project to navigate all data within CORD-19.

Target Tables

- Target tables contain the schema of how the medical literature is extracted to each scientific question. The summarised information will be read and assessed by the researchers in order to answer some of the key questions related to COVID-19.
- In addition target tables include a sample list of literature which is associated with some of the key questions. This sample list of labelled data will be the 'ground truth' for the **supervised learning** approach proposed by this project.

Solution Statement

This project proposes a Machine Learning technique to address the overflowing of un-mined COVID-19 related medical literature. More specifically, supervised learning models (i.e. SVM, Random Forest) will be trained to classify medical literature within unsolved scientific questions. The model will be using Natural Language Processing techniques to transform text based information into features.

With new cases of COVID-19 increasing everyday, the numbers of research literature will continue to increase until the key questions related to COVID-19 are resolved.

The final result of this project is to propose an automated software solution such that key questions related to COVID-19 can be addressed and updated on a daily basis. The quality of the machine learning model will be continuously validated and tested in order to address the latest up to date research of COVID-19.

Benchmark Model

A traditional search engine can be implemented using the same CORD-19 dataset as the proposed machine learning model, so that the results are comparable.

Based on user search results, a search engine scans through all the matches within its index, and then ranks and assembles the results that are most relevant to the question using custom built algorithms. Some of the key features include keyword, linking, location, and freshness of the data.

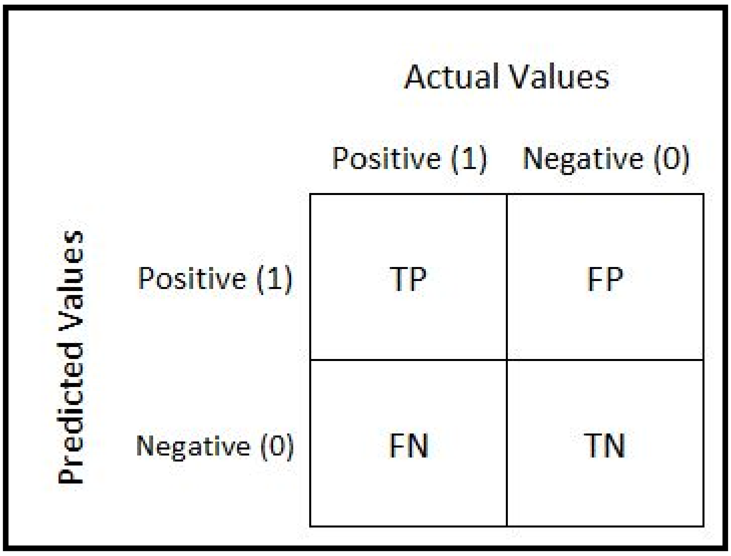
The native search engine might be able offer a suitable solution to classify and prioritise medical literatures into each key scientific problem. However, it may not be able to adapt to the rapid change of information during the pandemic.

The ability for machine learning models to learn continuously is beneficial in comparison to search engines. This project aims to build and train both models with the goal of comparing the outcome of the machine learning model versus a search engine. Then we will be able to evaluate whether the problem is better resolved with a machine learning method than to a search engine.

Evaluation Metrics

The quality of the models will be evaluated based on the same metrics. This enables us to quantify the comparison between the benchmark model and our proposed machine learning model. The following metrics

- **Precision -** Can the model correctly classify each medical literature into the correct questions without miss judging irrelevant literatures into the same question.
  - $Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$
- **Recall -** Can the model identify all the medical literatures that are relevant to the questions without missing any relevant literature.
  - $Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

[3]

Project Design

The theoretical design for implementing the proposed solution is described in the following:

1. Exploratory data analysis
    a. Perform a high level data visualisation in order to understand the structure and distribution of text data, including missing values and inconsistencies
    b. Perform structural analysis on the source file and the critical relationship between structures - i.e. title, text, and abstract.
    c. Understand the outcome and the target output of the model
2. Data Preparation and Cleansing
    a. Perform a fix to resolve issues found in exploratory data analysis.
    b. Distributing the data into Test, Train, and Validation sets Feature Engineering
3. Implementing Search Engine
    a. Creation of the benchmark model - this might be out of scope for this udacity course
4. Implement Machine Learning Algorithms
    a. Implement Natural Language Processing techniques such as bag of words, n-grams, entity recognition, TF-IDF (Term Frequency–Inverse Document Frequency) used to vectorize text data
    b. Creation of a supervised learning model using SVM and Random Forest
    c. Perform K-fold validation and optimisation techniques i.e. hyper parameter tuning to create the most optimal model
    d. Identify Bias and Variance in the model
5. Evaluation and Performance
    a. Compare performance based on the evaluation metrics to verify which solution performed better in solving the problem

[3] (n.d.). Understanding Confusion Matrix - Towards Data Science. Retrieved June 15, 2020, from https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

6. Deployment

a. Present a methodology on how to productionise and deploy model with on-going training and improvement of the model
b. Creating end-points that are accessible for researcher to access

References

Allen Institute For AI, 2020, COVID-19 Open Research Dataset Challenge (CORD-19), Available at: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge [Accessed 15/06/2020]

Office of Science and Technology Policy, 2020, Available at: https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/ [Accessed 15/06/2020]

Google, 2019, How Google Search Works, Available at: https://www.youtube.com/watch?v=0eKVizvYSUQ [Accessed 15/06/2020]

Sarang Narkhede, 2018, Understanding Confusion Matrix, Available at: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62 [Accessed 15/06/2020]

Google News, Coronavirus (COVID-19), Available at: https://news.google.com/covid19/map?hl=en-AU&mid=/m/0chghy&gl=AU&ceid=AU:en [Accessed 15/06/2020]

Silvio Mori Neto, 2020, Capstone Project Proposal, Available at: https://github.com/silviomori/udacity-machine-learning-capstone-starbucks/blob/master/proposal.pdf [Accessed 15/06/2020]

Moghazy, 2020, COVID-19 Literature Ranking + Web Scraping, Available at: https://www.kaggle.com/moghazy/covid-19-literature-ranking-web-scraping [Accessed 15/06/2020]