Name(s)

Anjo Pagdanganan

Project Number

S0824

Project Title

A Deep Learning Approach to E. coli Epidemic Prediction

Abstract

Objectives

E. coli ravaged the US in 2018, with the first outbreak tracked back to Yuma, Arizona infecting 210 and the second, tracked back to central California infecting 62. Outbreaks of this severity create paranoia around the produce at hand throughout the US. This leads to numerous agricultural industries coming under fire for an outbreak that they may not even have caused, such as the Salinas Valley. This damage is only dragged on considering that the CDC takes two to three weeks to recognize such an outbreak. Thus, this project applies a deep learning approach towards predicting trends in E. coli outbreaks to minimize the damage dealt to agricultural industries.

Methods

Using Python, I obtained search popularity data of phrases correlating with "e coli symptoms" from Google through the pytrends library. I also scraped approximately 13 years of weekly E. coli case data from the CDC with several web scraping libraries (namely BeautifulSoup and TQDM). I then used this data to train and evaluate an LSTM (Long-Short-Term-Memory) neural network with Keras. As a baseline metric, a naive forecast using the current week's data as predictions was used.

Results

RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) were used as accuracy metrics. The dataset obtained was split into 70% training data, 15% validation, and 15% testing, leaving approx. a year or two of testing data from each of the 9 CDC census regions. With an RMSE of 5.20 and MAE of 3.05 compared to the naive forecast's RMSE of 6.18 and MAE of 4.24, the model exceeded baseline performance, showing that it has predictive power. The model does have a slight flaw in that its predictions peaks around 15-25 cases, highlighting that it likely needs more data for a proper fit.

Conclusions

To my knowledge, this is the first time this technique has been used to predict food poisoning cases, although it was inspired by similar applications in dengue outbreak prediction. Despite being trained on a very small dataset, the LSTM neural network was able to recognize trends in E. coli epidemics. In the future, this model could potentially be developed into a dynamic E. coli epidemic prediction system, constantly tracking search queries and current E. coli data for training and future predictions.

Summary Statement

I applied forecasting techniques with deep learning towards predicting the number of E. coli cases in a specific region on a week-by-week basis.

Help Received

None. I researched all of the algorithms and scraped all the data by myself.