

Abstract

EMOS Master thesis competition 2023

‘Using Digital Trace Data to Generate Representative Estimates of Disease Prevalence [COVID-19 Infections] in Belgian Municipalities’

Author: Dishani Sen, KU Leuven

Keywords: COVID-19, Digital-Trace, Feature Extraction, Twitter, Bayesian Hierarchical Modelling, MrP

Is it possible to predict the area-level prevalence of COVID-19 infections in Belgium by analysing self-reported symptoms on Twitter? This research project is about generating estimates of the incidence of COVID-19 infections, at the municipality level, by using Multilevel Regression Post-Stratification (MrP) to account for sampling biases in the social media sample. At first, tweets are obtained from users based on keywords derived from previous research, e.g., tweets mentioning fever, cough, loss of taste, fatigue, etc. Then, key demographic and geographical features of interest are extracted using the M3 deep learning pipeline, as well as simple self-reported characteristics, effectively transforming the unstructured twitter sample into a survey-like object. Finally, based on these demographic features and census characteristics, a mixed effects logistic regression model with post-stratification according to the Belgian census is proposed to forecast the number of infected individuals on a particular day. This study intends to contribute to the proof of concept of a complete end to end pipeline to perform real time predictions of disease prevalence at a granular level in a population using social media data.