**Datasets:**

‘USA\_and Covid\_specific\_dataset.csv’

* The initial Reddit Comments dataset retrieved from SocialGrep.com, then filtered and pre-processed as specified in the methodology.

‘Time\_Series\_Reddit\_classified\_Dataset.csv’

* The generated time-series dataset, with features generated as specified in the methodology.
* All features except the target variables are normalised to having the same scale
* Below is a table describing each what each feature represents

|  |  |
| --- | --- |
| **Dataset Feature** | **Description** |
| Label: ('*YES'/'MAYBE'/'CONTACT'*) | Number of Reddit texts classified as labels;  Author: 'Yes' / 'Maybe', or Contact: 'Yes', grouped by day |
| {*Label*}\_Rolling\_Mean | Rolling average (n=7) of respective label |
| {*Label*}\_std | Rolling standard deviation (n=7) of respective label |
| {*Label}\_*diff | Rate of change between current and previous label value |
| {*Label*}\_Rolling\_Mean\_dfff | Rate of change between current and previous label (mean) value |
| {*Label*}\_std\_diff | Rate of change between current and previous label std |
| New\_cases | WHO reported Covid-19 incidence (Target Variable) |
| Previous\_New\_cases | Covid-19 incidence shifted by 1-day |
| 7\_day\_New\_cases | Covid-19 incidence lagged by 7 days (Target Variable) |
| 14\_day\_New\_cases | Covid-19 incidence lagged by 14 days (Target Variable) |