How the use of Twitter can assist in the prediction of the number of new cases of COVID-19 in British Columbia

1 Introduction

This study aims to predict the number of new infected people with COVID-19 at British Columbia (BC), Canada based on the number of issued tweets with the hashtag #COVID19 and on the use of some particular keywords in these tweets. The motivation for this paper comes from Hideo Hirose and Liangliang Wang's Prediction of infectious disease spread using twitter: A case of influenza [1], where the Ridge logistic regression produces the best predictions in the context of influenza.

2 Methods

2.1 Data

Information about the daily detected cases of COVID-19 at BC was collected from the "BC Centre for Disease Control". Additionally, we collected a sample of 124,335 tweets containing the hashtag #COVID19 that were issued by registered accounts with English language at BC and surrounding areas during the period from 13th April to 26th April of 2020 through the standard public API of Twitter.

2.2 Target

Linear regression models predict values in the range $(-\infty, \infty)$, but our target, which is the daily number of new COVID-19 cases in BC, belongs to the range $[0, \infty)$. Thus, we transform the target variable to $\log(y + 0.01)$, where y denotes the target variable, and the small quantity 0.01 is added to avoid getting an undefined number when there are zero cases. After fitting a model, the predicted number of new cases can be obtained by applying the exponential function to the transformed predicted value and rounding it to zero digits. In this way, linear regression models predict values in the range $[1, \infty)$.

2.3 Features

We propose two sets of features. The first one contains, in addition to the daily number of issued tweets, the daily percentage of tweets that contains a keyword. Each included keyword is determined by a similar criterion as proposed in [1], which consists in the main symptoms of COVID-19 according to [2] and [3]: fever, tiredness, dry cough, aches and pains, nasal congestion, runny nose, sore throat, diarrhoea, headache, shaking with chills, as well as loss of smell and taste.

The second set of features is an extended version of the first one. To build it up, we lemmatize and remove urls and stopwords from the contents of tweets in the training set; then we analyze them through the fit of a word2vec model with a window of size 3 and the same hyperparameters values used in [4], using the package `gensim` and the support of the tutorial in [4]. In this analysis, we identify words and bigrams that share the most similar contexts with the keywords that comprise the first set of features. Next, we increase the set of keywords by including the following terms: illness, ill, suspected, severe, complication, "stay home", "feel like", contagious, "cause death", irresponsible, sneeze, ambulance, "self isolation", breath, breathing, infect, infected, control, disease, immunity, antibody, immune, prevent, spread, "wear mask", respiratory , expose, exposed, distancing, surveillance and death. Then, we extend the first set of features by including the daily percentage of tweets containing these new keywords.

3 Experiments

3.1 Multiple Lasso regression

Lasso is a model that tends to prioritize solutions with fewer non-zero solutions due to L_1 regularization. The objective function in this model is:

$$f(w) = \frac{1}{2n}||Xw - y||^2 + \alpha||w||_1,$$

where n represents the number of examples, α is a constant, and $||w||_1$ is the L_1-norm of the vector w.

3.2 Multiple Ridge regression

Ridge is a model that uses L_2 regularization. The objective function in this model is:

$$f(w) = ||Xw - y||^2 + \alpha||w||^2.$$

3.3 Robust regression (Hubber loss)

This model uses the Hubber loss, making it less sensitive to the presence of outliers, such as the number of of new cases observed on the 21st and 25th of April (68 and 95 cases, respectively). The objective function in this model (Hubber loss) is:

$$f(w, \sigma) = \sum_{i=1}^{n} \left(\sigma + H_\epsilon \left(\frac{X_i w - y_i}{\sigma} \right) \sigma \right) + \alpha||w||^2,$$

where

$$H_\epsilon(z) = \begin{cases} z^2, & \text{if } |z| < \epsilon \\ 2\epsilon|z| - \epsilon^2, & \text{otherwise} \end{cases}$$

We minimize this loss function over w and σ. As advised in the scikitlearn documentation, we set the parameter epsilon to 1.35 to achieve 95% statistical efficiency.

3.4 Random forest

We give no limit for the depth of the tree.

3.5 Cross validation

We use 5-fold cross validation to tune the hyperparameters corresponding to the regularization parameter α and the number of trees in linear regression and random forest models, respectively. The chosen hyperparameter values are given in tables 2 and 3.

4 Results

In order to determine if the prediction of the proposed models improves with the addition of features related to the content of tweets we fitted a simple linear regression and a random forest model, where the unique feature was the daily number of issued tweets with the hashtag #COVID19; the training and validation mean square errors are presented in Table 1. In Table 2 and Table 3 we present the results obtained with our models for the first and second sets of features, respectively.

Model	Training MSE	Validation MSE
Single linear regression	0.2115	0.4575
Random forest regression (100 trees)	0.0649	0.3257

Table 1: Average mean square error obtained by models using the daily number of issued tweets as unique feature.

Model	Training MSE	Validation MSE
Multiple linear regression	1.9753	0.3409
Multiple LASSO regression ($\alpha = 2.75$)	0.2244	0.3169
Multiple Ridge regression ($\alpha = 0.25$)	0.0783	0.1618
Multiple Robust regression ($\alpha = 0.25$)	0.0682	0.2003
Random forest regression (50 trees)	0.0293	0.1873

Table 2: Average mean square error obtained by each model with the training and validation set.

Model	Training MSE	Validation MSE
Multiple linear regression	0.0000	0.5178
Multiple LASSO regression ($\alpha = 0.75$)	0.2259	0.2478
Multiple Ridge regression ($\alpha = 3.75$)	0.0849	0.2537
Multiple Robust regression ($\alpha = 1.75$)	0.1752	2.0402
Random forest regression (800 trees)	0.0327	0.2217

Table 3: Average mean square error obtained by each model with the training and validation set.

4.1 Test error

After analyzing the results of our experiments, we picked the multiple Ridge regression which was trained with the first set of features and we calculated the test error, which corresponds to an MSE of 0.0429. The predicted number of cases (32) is shown in the following Figure, where it is marked with the redpoint.

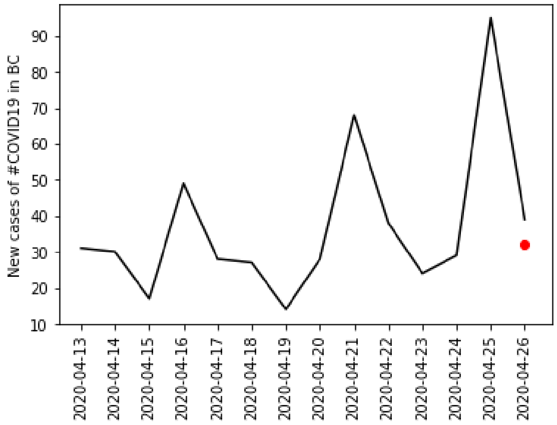

Figure 1: The observed number of new daily cases of COVID-19 from 13th April to 26th April at BC, and the predicted number of new cases in 26th April according to our model.

5 Conclusion

We draw attention to the low validation and test errors, particularly with the multiple Ridge regression model using the first data set. Therefore, despite the limited amount of data and the omission of the temporal structure of the data, which involves an inherent correlation between the number of cases per day, this study shows it is possible to obtain information on the COVID-19 spread from Twitter data, more specifically from the number of issued tweets with #COVID19 and the inclusion of the first set of keywords in their content, which is mainly comprised of symptoms. Finally, we add that it be may possible to extend this study to other types of social media, provided there is access to enough data.

References

[1] Hideo Hirose and Liangliang Wang. Prediction of infectious disease spread using twitter: A case of influenza. In *2012 Fifth International Symposium on Parallel Architectures, Algorithms and Programming*, pages 100–105. IEEE, 2012.

[2] https://www.cdc.gov/coronavirus/2019-ncov/symptoms testing/symptoms.html.

[3] https://www.google.com/search?q=covid+19+symptoms&oq=COVID19+sym&aqs=chrome.

[4] https://www.kaggle.com/pierremegret/gensim-word2vec tutorial.