Natural Language Processing to Analyze the Effectiveness of Social Media on Spread of COVID-19

Abhinav Mandava
B00841453
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia
Email: Abhinav.Mandava@dal.ca

Anqi Chen
B00838586
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia
Email: Anqi.Chen@dal.ca

Ahsan Kamal
B00853723
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia
Email: ah721730@dal.ca

Ram Prasath Meganathan
B00851418
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia
Email: rm507817@dal.ca

Raj Kumar Reddy Gangis
B00849566
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia
Email: Raj.Gangi@dal.ca

***Abstract*—Social Media becomes an effective way for social influencers to directly raise public attention to act on important things. In North American, Twitter is one of the most influential digital information media. During Covid-19, the change in cases is an important index to see whether public health authorities' social media behaviour impacts the public's positive response. In short, this study intends to analyze how long the possible influence of tweets may take on the change in daily cases.**

1. Introduction

Everyone around the world has been affected by COVID-19, since the end of last year. There are various data changes about the epidemic every day. At the same time, media reports and dissemination also have a certain impact on citizens' social activities, which may lead to the differences in the influence of Coronavirus. In this assignment, we are analyzing the effectiveness of social media (Twitter), the decisions and information from authoritative institutions and health organizations, on the spread of COVID-19.

2. Methods

2.1. Data Resources

2.1.1. Google Big query. We fetched the first dataset from Google Big-query public dataset [1]. The region is specified as Canada and the date range is between November 1, 2019 to September 1, 2020.

2.1.2. Tweets from Three Accounts. Three twitter accounts we selected are @CanadianPM, @Safety_Canada, @CDCgov. The date range considered is from November 1, 2019 to September 1, 2020. There're 3738 tweets collected for those days.[2]

2.2. Preprocessing Data

First we combine tweets from three sources into a cell of text content. The table now only contains 355 rows of data with date and combined tweets. Then calculate the case changes, put the corresponding labels (increase, no change, decrease) for each day. If the current days' new cases are more than the previous day, we labelled that record as "Increase the spread of COVID19", if less than the last day, we labelled as "Decrease the spread of COVID19" else "No change on the spread of COVID19." There are around 100 records with no change, 112 records with increase in spread and 121 records with decrease in spread of COVID19. Lastly classify data for two models with different time slots, one is 1 day and the other is 14 days.

2.3. Use Word2Vec to Vectorize Tweets [3]

After all the tweets are read, they are tokenized into words and are separated into three different data sets. These three different datasets were then passed to the NLTK library which is a toolkit for removing the stop words from the tweets in order to prevent vectorizing those stop words.

These three different tweets are fed into the Word2Vec algorithm which is a natural language processing algorithm for reading the text and vectorizing them into numerical values. The vectorized values can then be fed to the machine learning models for further classification.

We have also researched the use of other word preprocessing algorithms such that Glove and found that Word2Vec

is widely used when it comes to capturing semantic similarities.

Here we clustered the tweets in to three clusters, we have used K-means clustering algorithm from NLTK and Scikit-learn libraries to find the boundaries between the tweets.

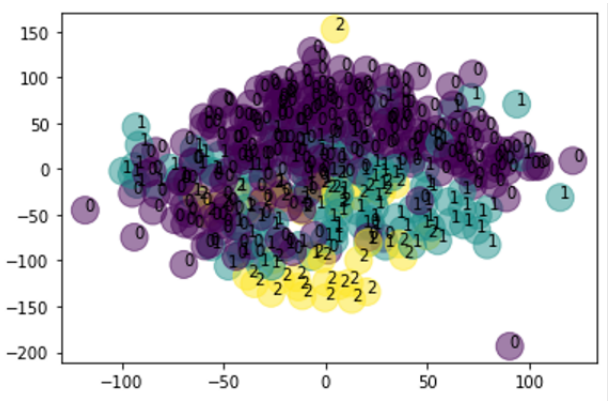

Fig1.Clustering Tweets[4]

2.4. The K-Nearest Neighbour Classifier

K-Nearest Neighbour classifier is a widely used classification algorithm, especially for multiclass classification. The KNN classifier takes the input feature set X and finds the mapping to the class label Y_i for each feature vector X_i in X. KNN uses a closeness/distance metric to find the K-nearest neighbouring vectors for the given feature vector X_i. The class label for this feature vector X_i would then be the same as that of its K-nearest neighbours.

In our context, each feature vector X_i corresponds to a tweet which was vectorized using the Word2Vec word embedding technique. The model was trained on the tweets taken from November 2019 to September 2020. The tweets from April 2020 to September 2020 have been used to see the model's performance by tweaking the hyperparameters.

2.4.1. Hyperparameter Tuning. In any machine learning algorithm, there are several algorithmic parameters which when tweaked give results with varying accuracy. The goal is to search for a good set of such hyperparameter values which would maximize the accuracy and minimize overfitting. However, finding such hyperparameter values is not easy, as that would require searching for optimal solution in a very large search space. There are several methods for tuning the hyperparameters and is one of the important areas in machine learning, subject to research. One such method is Grid Search which we used in our model.

Grid Search for Hyperparameter Optimization: Grid search, or parameter sweep is a traditional method[5] to find the best combination of hyperparameter values, by taking the set of values for each of the hyperparameters from the user. When we give the desired values of each hyperparameter, then the search space can be abstracted as a high-dimensional grid. A grid search is driven by a performance metric[5], which is typically a cross-validation score. The combination of hyperparameter values with the best score gets to be the best-params set for the model. For our KNN model, the hyperparameters considered were n_neighbors, weights, and metric.

The drawback of grid search is that it is an exhaustive search which can require high computational power or time when we consider a lot of hyperparameters. Some other examples of hyperparameter learning techniques include Bayesian Optimization, Random Search, Gradient-based Optimization, Genetic Algorithms (Evolutionary Algorithms), etc. [5]

2.4.2. Model Evaluation. The table below shows the training accuracy and testing accuracy for both Model A and Model B.

Models	Model A	Model B
Training Accuracy	87.18%	98.47%
Testing Accuracy	39%	50%

Confusion Matrices:

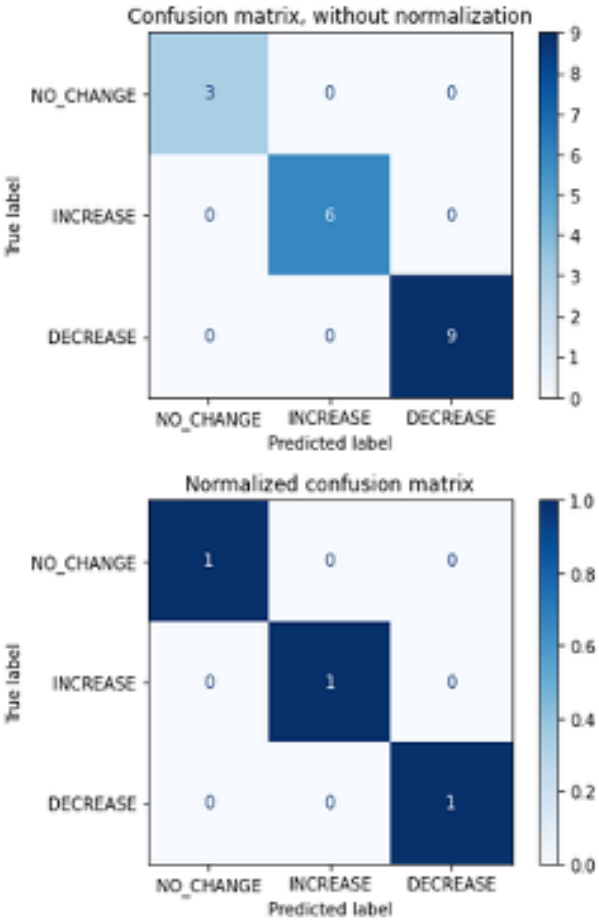

Fig2A.Confusion Matrix for Model A

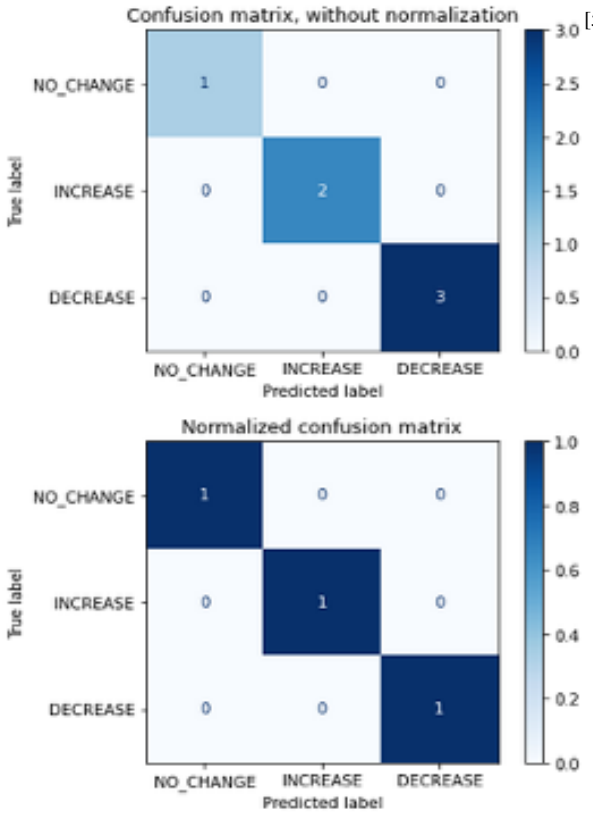

Fig2B.Confusion Matrix for Model B

3. Conclusion

Compared these two models (Model A and Model B), we suggest that model B should be used to evaluate a message published by influencers because the accuracy for both training and testing is higher than the other one. The training accuracies respectively are 87.18% and 98.47%, testing accuracies respectively are 39% and 50%,which can also be found in notebook.

References

[1] "BigQuery public datasets — Google Cloud", Google Cloud, 2020. [Online]. Available: https://cloud.google.com/bigquery/public-data. [Accessed: 19-Oct-2020].

[2] "A script to download all of a user's tweets into a csv", Gist, 2020. [Online]. Available: https://gist.github.com/yanofsky/5436496. [Accessed: 19- Oct- 2020].

[3] "Text classification using word2vec", Kaggle.com, 2020. [Online]. Available: https://www.kaggle.com/ananyabioinfo/text-classification-using-word2vec. [Accessed: 19-Oct-2020].

[4] "K Means Clustering Example with Word2Vec in Data Mining or Machine Learning - Text Analytics Techniques", Text Analytics Techniques, 2020. [Online]. Available: https://ai.intelligentonlinetools.com/ml/k-means-clustering-example-word2vec/. [Accessed: 19-Oct-2020].

[5] "Hyperparameter optimization", En.wikipedia.org, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Hyperparameter_optimizationGrid_search. [Accessed: 19- Oct- 2020].