**Topic**

Sentiment Analysis of Change in Working Style due to Coronavirus Outbreak

**Project Members**

Tan Sia Hong

Teo Boon Long

Tarashini A/P Suthesan

Liu Hong Yang

**Problem Statement**

The rapid spread of the Coronavirus pandemic has a huge impact on daily lifestyles in terms of fundamental lifestyle, learning style, working style and leisure lifestyle. It is learnt 81% of 3.3 billion global workforce has been affected due to complete or partial workplace during the outbreak. International companies like Google, Microsoft, Twitter, Hitachi, Apple, Amazon, Chevron, Salesforce, introduced obligatory work-from-home as their business strategies. However, the reality of working remotely during lockdown has its own set of challenges like inadequate communication, increased work hours and screened time and mental health. poor connectivity issue, inconducive workspace at home.

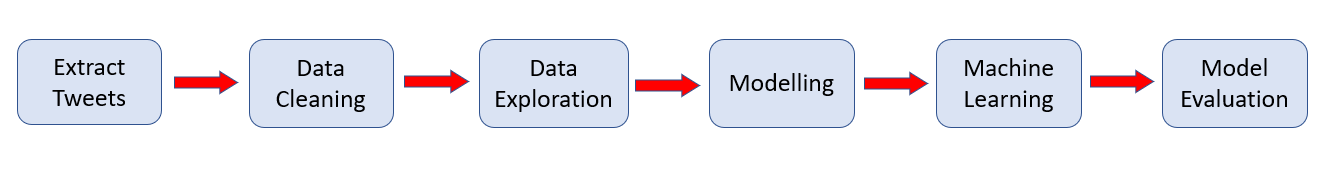
**Aim**

This project focuses to investigate on the impact of working from home due to coronavirus outbreak

**Objectives:**

1. To identify positive & negative sentiments towards working from home revolution
2. To train the models to identify working from home revolution perception using different machine learning approaches (Naive Bayes, Decision Tree, Logistic Regression)
3. To evaluate performance of the models

**Methodology:**

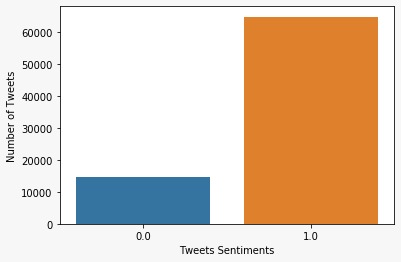
****

1. Gather Data (Tweets Extraction)

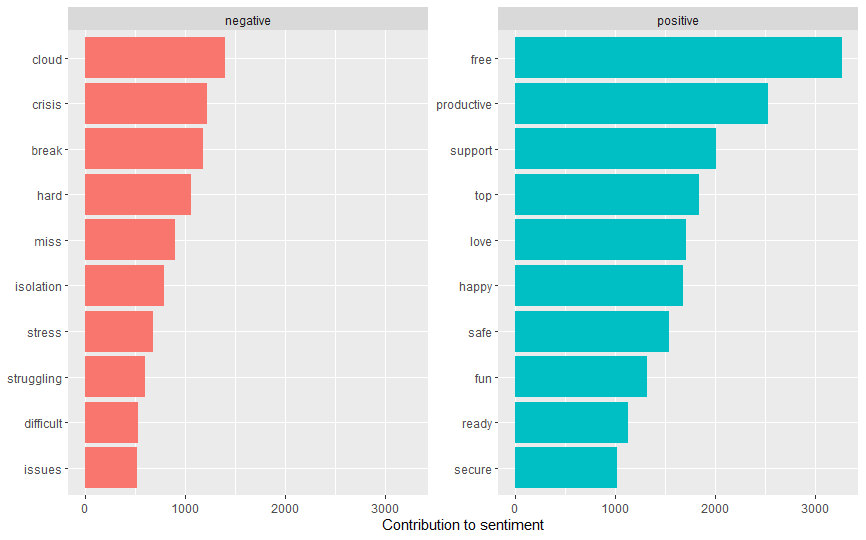
Extracted in total of 115,000 Tweets during the month of April using **#wfh**

1. Data Preparation (Pre-processing and Cleaning Tweets)
   1. Word Tokenization
   2. Remove stop words, emojis, special characters, and extra blank spaces
   3. Stemming
   4. Lemmatization
2. Data Exploration

In the dataset, polarity of the tweets ranges from 0 to 1, with 0 being negative and 1 being positive and was categorized using Python Vader library. 35,645 out of 115,000 tweets were excluded as those tweets had NULL value. As can be seen in the bar plot below, out of 79, 355 tweets, there are 64838 positive and 14517 negative tweets respectively.



The dataset is further explored in terms of word counts to compare positive and negative tweets. Based on the following bar plot, most used positive words in the extracted tweets are free, productive, support, top, love, happy, safe, fun, ready and secure. On the other hand, negative tweets mostly contain words like cloud, crises, break, hard, miss, isolation, stress, struggling, difficult and issues. Overall, we found that 80% of the tweets related to working from home during the pandemic is positive. Hence, we could generalize that impact of working from home due to coronavirus outbreak is indeed perceived positively by majority.



**Modelling (Feature Engineering)**

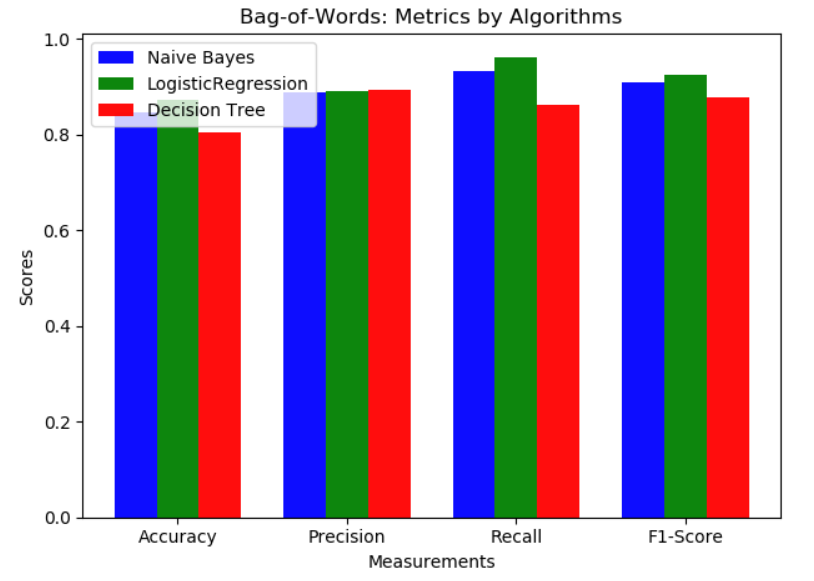
Next, the tweets are translated into numerical format that’s easily readable by the machine. The two types of feature engineering algorithms used for translation in this project are Bag-of-words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF). BOW creates a set of vectors with the count of word occurrences, while TF-IDF contains information on more and less important words too.

**Machine Learning**

The project used 3 supervised techniques for sentiment classification which are Naïve Bayes, Logistic Regression and Decision Tree. *Naïve Bayes*, supervised classifier, uses conditional probability to classify words into positive and negative sentiments. *Logistic Regression* is another supervised classifier which can be relied on for binary classification problems, in this project are namely positive and negative tweets. *Decision tree*, supervised classifier which is trained on annotated data and calculates how much each tweet correlates with a particular label. The data in this project is split into 70% of training set and 30% of testing set to measure the accuracy of the models.

Model Evaluation

Comparison of Naïve Bayes, Logistic Regression and Decision Tree performance using Bag-Of-Words as feature engineering technique can be observed below.

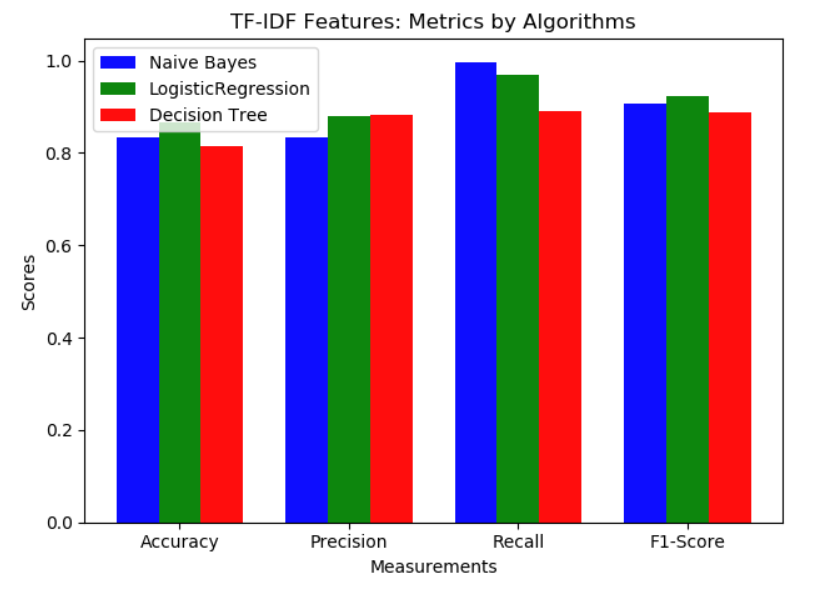


>>>explain

We going to use the bag of word approach to classify. In bag of word approach, we will take all the word in every text from tweets data then count the number of occurrences of each word. After finding the number of occurrences of each word, we will choose certain number of words that appeared more often than other words. The project is used three model for classification and show their performance as above graph. The three-classification algorithm that work with text data which is Naïve Bayes, Decision tree and logistic regression. Based on the graph above, their accurate is achieved around score of 0.8 but we have obtained an accuracy of the highest is logistic regression compare with two others. Therefore, we can conclude the best model to predict the class for test set is logistic regression.

>>Add on

Comparison of Naïve Bayes, Logistic Regression and Decision Tree performance using TD-IDF as feature engineering technique can be observed below.



>>>explain

The three algorithms to predict with text class is increased the accuracy using the TF-IDF feature. However, the logistic regression algorithm had highest accuracy score compare with other so was chosen for prediction model.

>> Evaluation on algorithms

To evaluate the performance of the algorithms, we use different metrics to measure the performance of the algorithms. Multiple metrics combined together could avoid the failure of classification that the problem that sole measurement causes. In the older version, we just utilize the single measurement: accuracy. While, the accuracy is not the suitable measure for assessing classification classes. As it may be a poor measurement for the imbalance data. Then we add another three metrics: Precision, Recall and F1-Score. The F1-Score could combine the procession and recall to give better evaluation results. From the Bag-of-Words and TF-IDF methods, we could notice that Logistic Regression algorithm behaves better any the another two algorithms. In further work, we could modify its and then find the best sets of parameters which could apply to unlabeled data.