**Tweet Analysis of the Coronavirus**

Abhishek Patil

**ABSTRACT**

The examination of data to identify feelings using algorithms that allow us to ascertain if people have positive or negative emotions regarding a topic is a fundamental challenge in machine learning. Social networks and microblogging, which are mostly used to share personal opinions and thoughts, are excellent sources of information. Based on this information, we suggest analysing the sentiment of English tweets posted during the COVID-19 epidemic in 2020. By utilising the Logistic Regression model, we were able to categorise the tweets as positive or negative with an accuracy of almost 90%.

**PROBLEM STATEMENT**

The pandemic illnesses that are currently plaguing the world, in particular, pose major issues for the population on all fronts: economic, emotional, status, planning, and politics, in addition to the complexity of customs, ethics, personal psychology, and social behaviour of individuals. Therefore, when difficult situations happen, it is essential and required to analyse people's views. Observing how people respond to this threat can reveal crucial details about how society behaves and responds to unforeseen events, whether positively or negatively. At the moment, the internet and social media have evolved into potent tools for gathering people's opinions and comments on a variety of subjects.

The major goal is to create a prediction model that could aid in anticipating a tweet's sentiment.

**ATTRIBUTE INFORMATION**

* Location
* Tweet at
* Original Tweet
* Sentiments

**INTRODUCTION**

The respiratory infections caused by human respiratory viruses are among the most prevalent viral illnesses that affect people (RVs) The influenza virus, also known as the "flu," is the most well-known type of respiratory viral infection and accounts for between 250,000 and 500,000 fatalities annually globally. The H1N1 virus is its most well-known variety. The corona virus is one of the virus families that cause respiratory illnesses. It infects the epithelial cells of the respiratory tract in humans, frequently going undetected but occasionally being fatal. The Middle East Respiratory Syndrome (MERS), severe acute respiratory syndrome (SARS), and currently Corona virus Disease are the most well-known corona virus varieties (COVID-19).

In the middle of December 2019, the first cases of people exhibiting respiratory sickness brought on by the corona virus were reported. The Wuhan Health Commission in the Chinese city of Wuhan revealed data on cases of atypical pneumonia affecting patients who came from a neighbourhood market on December 31, 2019. Outside of China, more than 4500 cases and more than 60 fatalities associated with COVID-19 had been confirmed by late February 2020. The World Health Organization (WHO) declared the COVID-19 disease to be a pandemic on March 11, 2020, when it was estimated that 118,000 people were afflicted in 114 countries, and 4,291 deaths had been confirmed. This was done in response to the corona virus's worrisome levels of severity and transmission.

The seriousness of the virus was discovered by the first COVID-19 impact analysis on humans. 3 (4.5%) of the 67 patients had light illness, 35 (52.2%) had moderate illness, 22 (32.8%) had severe illness, and 7 (10.4%) had critical illness. The computer tomography method was employed to assess the disease's severity level (CT)

A key component of understanding how people perceive the COVID-19 pandemic's effects on social, economic, political, and technological levels is the application of machine learning algorithms for the search of emotions represented in text on microblogging social networks. The following article is based on a monthly analysis of English tweets since the outbreak started.

**STEPS INVOLVED**

In order to go ahead for data visualization upon key factors we need to go for certain extra steps before proceeding to the main segment. In this part we are going with the following steps:

1. Importing Analytical necessary library classes for future analysis.
2. Reading the csv data file from Google drive.
3. Setting figure size for future visualization.
4. Removing future warnings in seaborne plots.
5. Visualizing all the columns of the respective Data frame.
6. Viewing all data information
7. Checking the Unique values in the column ( if any)
8. Converting the data types to similar objects as the Analysis Demands.
9. Formatting the “size” column into a single column in the dataset.
10. Eradicating special characters from the dataset columns.

* **EXPLORATORY DATA ANALYSIS**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypotheses and to check assumptions with the help of summary statistics and graphical representations. It gives us better idea of which feature behaves in which manner compared to target variable. After loading the dataset we performed this method by comparing our target variable that is Rented Bike count with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables.

* **EXAMINING** **NULL** VALUES

The most critical thing from which we can draw some observations is Dataset, however data comes with unexpected values too i.e. sometimes it may be Null or missing in other words the space might be blank. Thus, at the time of analysing the first thing which we will do is to examine the null or missing values on the Dataset. It is the first step that will make the results “more” accurate &should be handled before it affects the performance of the models that predict the outcome.

* **NATURAL LANGUAGE PROCESSING**

The difference between text mining and natural language processing must be taken into account. Text mining focuses on the discovery and extraction of information of interest within an unstructured text, whereas natural language processing aims to perform an extraction of a more complete meaning indicator from a text, trying to find out who, when, where, how and why an action was performed, to achieve it, NLP applies complex algorithms to perform different types of analysis such as morphological and lexical, syntactic, semantic, discourse integration and pragmatic

* **Tokenizing**

TF = number of times a word appears in a document / total number of words in the document

*n*

The process of tokenization is the division of a long text into sentences which will be delimited by punctuation marks showing the end, or by words creating a list that stores all the words in a text individually.

* **STEMMING**

IDF = total number of documents / number of documents where a word is present

Stemming is the technique that identifies conjugated words and represents them in a unique way that expresses the same meaning and works with heuristics that seeks to cut out the words to standardize all conjugations and derivations. On the other hand, the Lemmatization technique applies a more complex analysis that through a word morphological analysis tries to find the base form of the conjugated words e.g. "am", "are", "is" is represented with its base form "be"

* **TF-IDF**

(Term Frequency - Inverse Document Frequency), this is a technique to quantify a word within a text, thus weights will be assigned to each one of the words, which means the importance of the word in the document, and is calculated as follows:

TF-IDF = Term Frequency (TF) \* Inverse Document Frequency (IDF)

Where TF is the frequency of a word in a document, while IDF is the inverse of the number of documents where the word is present.

Each document and word has its own TF, and is given by:

The IDF measures the informativeness of a word, or how rare is to find a word in a certain number of documents:

In case a word is not contained in any document, DF will be equal to 0, as dividing by 0 is undefined, the last equation was modified, having:

IDF = log(total number of documents /(number of documents where a word is present + 1))

IDF = log(N/(DF+1))

Finally, combining the equations, it is possible to get the TF-IDF score:

**TF-IDF= (t/n) \* log(N/(DF + 1))**

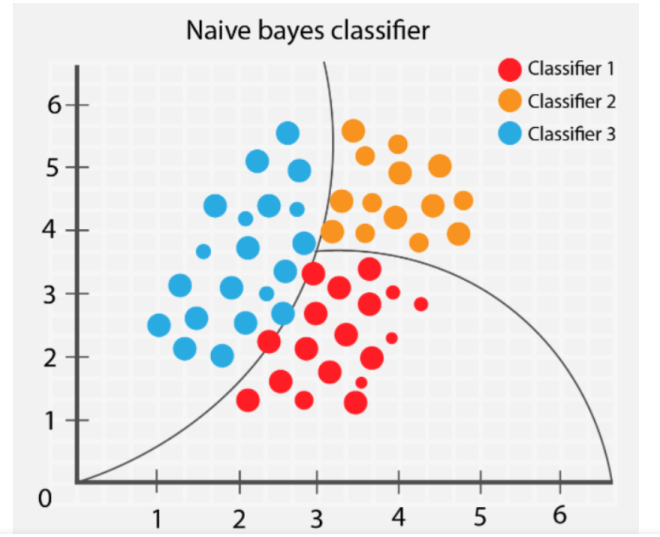
**FITTING DIFFERENT MODELS**

* Naive Bayes
* Stochastic Gradient Descent (SGD)
* Random Forest
* Support Vector Machine
* Logistic Regression

**ALGORITHMS**

# I. Naive Bayes

The Naive Bayes classification algorithm is a probabilistic classifier. It is based on probability models that incorporate strong independence assumptions. The independence assumptions often do not have an impact on reality. Therefore they are considered as naive. You can derive probability models by using Bayes' theorem (credited to Thomas Bayes). Depending on the nature of the probability model, you can train the Naive Bayes algorithm in a supervised learning setting.

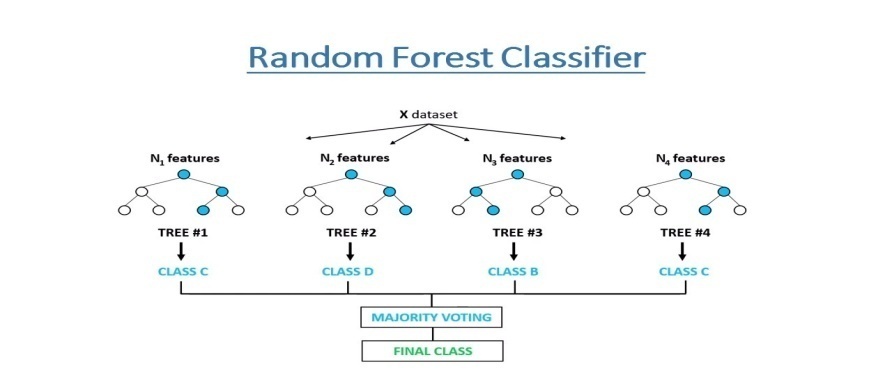


**II. Stochastic Gradient Descent**

Stochastic Gradient Descent (SGD) is a simple yet efficient optimization algorithm used to find the values of parameters/coefficients of functions that minimize a cost function. In other words, it is used for discriminative learning of linear classifiers under convex loss functions such as SVM and Logistic regression. It has been successfully applied to large-scale datasets because the update to the coefficients is performed for each training instance, rather than at the end of instances.

**III. Random Forest Classifier**

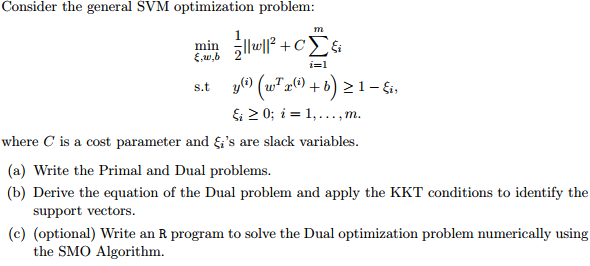
Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly selected subset of the training set, collects the labels from these subsets and then averages the final prediction depending on the most number of times a label has been predicted out of all.

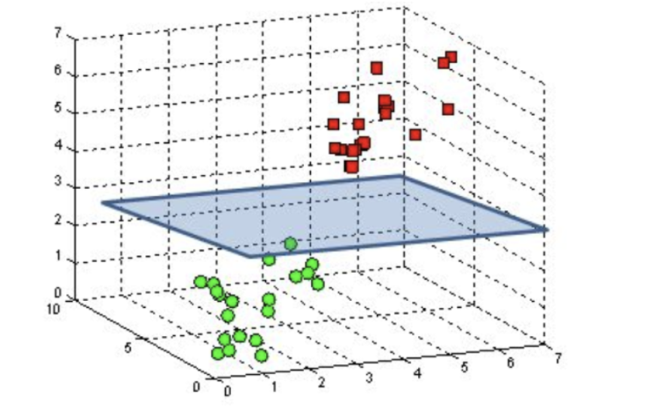


**IV. Support Vector Machine Classifier**

SVM is used mostly when the data cannot be linearly separated by logistic regression and the data has noise. This can be done by separating the data with a hyperplane at a higher order dimension.

In SVM we use the optimization algorithm as:





We use hinge loss to deal with the noise when the data isn’t linearly separable.

Kernel functions can be used to map data to higher dimensions when there is inherent nonlinearity.

**Logistic Regression**

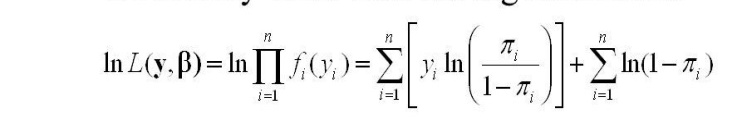
Logistic Regression is actually a classification algorithm that was given the name regression due to the fact that the mathematical formulation is very similar to linear regression.

The function used in Logistic Regression is sigmoid function or the logistic function given by:

f(x)= 1/1+e ^(-x)



The optimization algorithm used is: Maximum Log Likelihood. We mostly take log likelihood in Logistic:



**MODEL PERFORMANCE**

**Confusion Matrix-**

The confusion matrix is a table that summarizes how successful the classification model is at predicting examples belonging to various classes. One axis of the confusion matrix is the label that the model predicted, and the other axis is the actual label.

**Precision/Recall-**

Precision is the ratio of correct positive predictions to the overall number of positive predictions: TP/TP+FP

Recall is the ratio of correct positive predictions to the overall number of positive examples in the set: TP/FN+TP

**Accuracy-**

Accuracy is given by the number of correctly classified examples divided by the total number of classified examples. In terms of the confusion matrix, it is given by: TP+TN/TP+TN+FP+FN

**Area under ROC Curve (AUC)-**

ROC curves use a combination of the true positive rate (the proportion of positive examples predicted correctly, defined exactly as recall) and false positive rate (the proportion of negative examples predicted incorrectly) to build up a summary picture of the classification performance.

**HYPER PARAMETER TUNING**

Hyperparameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyperparameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

We used Grid Search CV, Randomized Search CV and Bayesian Optimization for hyperparameter tuning. This also results in cross validation and in our case, we divided the dataset into different folds. The best performance improvement among the three was by Bayesian Optimization.

**Grid Search CV**

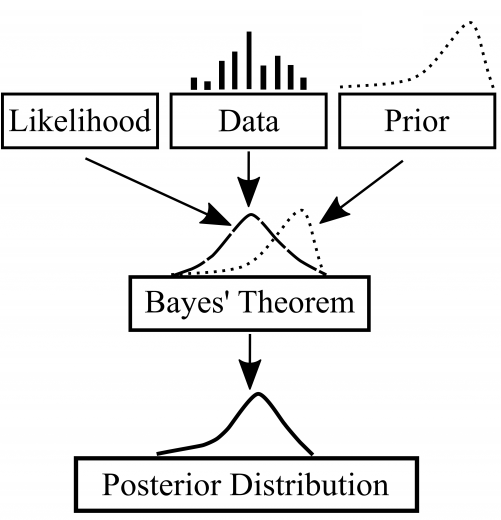
Grid Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model’s performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.

**Randomized Search CV**

In Random Search, the hyperparameters are chosen at random within a range of values that it can assume. The advantage of this method is that there is a greater chance of finding regions of the cost minimization space with more suitable hyperparameters, since the choice for each iteration is random. The disadvantage of this method is that the combination of hyperparameters is beyond the scientist’s control

**Bayesian Optimization**

Bayesian hyper parameters optimization is a very efficient and interesting way to find good hyper parameters. In this approach, in naive interpretation way is to use a support model to find the best hyper parameters. A hyper parameter optimization process based on a probabilistic model, often Gaussian Process, will be used to find data from data observed in the later distribution of the performance of the given models or set of tested hyper parameters.



As it is a Bayesian process each iteration, the distribution of the model’s performance in relation to the hyperparameters used is evaluated and a new probability distribution is generated. With this distribution it is possible to make a more appropriate choice of the set of values that we will use so that our algorithm learns in the best possible way.

**CONCLUSION**

The initial COVID-19 impact on people investigation revealed the virus's seriousness. 3 (4.5%), 35 (52.2%), 22 (32.8%), and 7 (10.4%) of the 67 patients had mild, moderate, or severe illnesses, respectively. The severity of the disease was evaluated using the computer tomography technique (CT)

The use of machine learning algorithms for the search of emotions reflected in text on microblogging social networks is a crucial part of understanding how people experience the effects of the COVID-19 pandemic on social, economic, political, and technological levels. The subsequent essay is based on a recurring study of English-language tweets sent out since the outbreak began.