



# USING TWITTER DATA TO IDENTIFY PUBLIC SENTIMENT TOWARDS THE COVID-19 PANDEMIC IN IRELAND

FINAL REPORT

HIGHER DIPLOMA IN SCIENCE IN DATA ANALYTICS

Words: 4809

SALIL RAINA

10527250@mydbs.ie

25/09/2020

# CONTENTS

1	INTRODUCTION.....	4
1.1	Novel Coronavirus.....	4
1.2	Sentiment Analysis.....	5
1.3	How does this work?.....	5
2	METHODOLOGY.....	8
2.1	Introduction to CRISP-DM.....	8
2.2	COVID-19 tweets sentiment analysis.....	10
2.2.1	Business Understanding.....	10
2.2.2	Data Understanding.....	11
2.2.3	Data Preparation.....	13
2.2.4	Data Modelling and Evaluation.....	20
2.2.5	Deployment.....	25
3	RESULTS.....	25
4	CONCLUSION.....	29
5	APPENDIX.....	31
5.1	Appendix 1 – Python script for Data collection.....	31
5.2	Appendix 2 – R script to merge different .csv files.....	32
5.3	Appendix 3 – Cleaning dataset and performing Sentimental analysis in R.....	33
5.4	Appendix 4 – Word Cloud in R.....	36
6	REFERENCES.....	39

## TABLE OF FIGURES

Figure 1. Lexicon based Sentiment Analysis approach .....	6
Figure 2. Machine learning classifier implementation .....	7
Figure 3. CRISP-DM Methodology .....	8
Figure 4. Snapshot of data from the Limerick.csv file .....	12
Figure 5. New Column 'Location' created in the .csv file.....	13
Figure 6. Final transformed dataset .....	14
Figure 7. Final cleaned and labelled dataset.....	16
Figure 8. Process in RapidMiner for sampling and balancing the dataset.....	17
Figure 9. Example of Vectorization done in Rapidminer .....	18
Figure 10. Example of 5-fold cross validation .....	19
Figure 11. Selecting the task and column to predict values .....	20
Figure 12. Selecting the input variables.....	21
Figure 13. Selecting the recommended models in RapidMiner.....	21
Figure 14. Accuracy for each type of model .....	22
Figure 15. Classification error for each type of model .....	22
Figure 16. Table of results of the Rapidminer automodel.....	23
Figure 17. Tweets by Location and Mood.....	26
Figure 18. No. of tweets by Date and Mood.....	27
Figure 19. Tweets by Mood and Time .....	28
Figure 20. Covid-19 word cloud.....	29

# 1 INTRODUCTION

About 90% of the people in the world have begun sharing their thoughts and perspectives on a daily basis on various micro-blogging sites since it's an easy way to express your views in a short and simple manner.(Saini *et al.*, 2019) On these micro-blogging websites, people are able to post their opinions in real-time on a variety of topics, discuss issues and express a negative or positive sentiment on a particular product or services they use or have used.(Agarwal *et al.*) The sentiments can be used to study user reactions about a product, service, event etc and then summarize an overall sentiment.(Agarwal *et al.*)

## 1.1 Novel Coronavirus

The Novel coronavirus disease (COVID-19) was initially detected in Wuhan, China, where a cluster of cases of pneumonia were reported. The disease, now known as COVID-19 spread to almost all parts of the globe and became a major public health concern for the countries.(*WHO Timeline - COVID-19, 2020*) Since December 2019, there has been a constant wave of infected cases and the outbreak now has been termed as a Pandemic. As of 21/06/2020, a total of 9,005,326 cases have been confirmed with 468,172 total deaths worldwide.(Worldometer)

The COVID-19 pandemic has been one of the most talked about topics currently over social media. With internet being accessible to a lot of people around the world, people have now begun to express their views over the internet.

## 1.2 Sentiment Analysis

For purpose of this project, we will be performing a sentiment analysis on twitter COVID-19 data. Firstly, it is important to understand what sentiment analysis is. Sentiment analysis is the interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques. Sentiment analysis uses natural language processing which allows businesses to identify customer sentiment toward products, brands or services in online conversations and feedback.(Learn, 2020) For example, by using sentiment analysis and analyzing 'n' number of tweets about a service would help a company to find if customers are happy or not.

The sentiment analysis is carried out using various techniques, such as Natural language processing (NLP), statistics and Machine learning (ML) approach. There are two main types of sentiment analysis, which are: Subjectivity/ objectivity identification and feature/ aspect-based sentiment analysis.(Nuggets, 2015)

- *Subject/ Objectivity Identification:* It consists of classifying a sentence or fragment of text into subjective or objective. There are challenges associated with this type of analysis as the sense of the word or even a phrase is often dependent on its context.
- *Feature/ Aspect-Based Identification:* It aims to determine the different sentiments and views in relation to various aspects of an entity. It allows more degree of overview of opinions and sentiments.

## 1.3 How does this work?

As mentioned above, Sentiment analysis uses various Natural Language Processing (NLP) methods and algorithms. Main types of algorithms used are as follows:(Learn, 2020)

- *Rule-based Approaches*

This type of systems performs sentiment analysis based on a set of manually determined rules which help in identifying polarity, subjectivity, or subject of a sentiment. Some techniques are 1. Stemming, tokenization, part-of-speech tagging, and parsing, 2. Lexicons (lists of words and expressions).

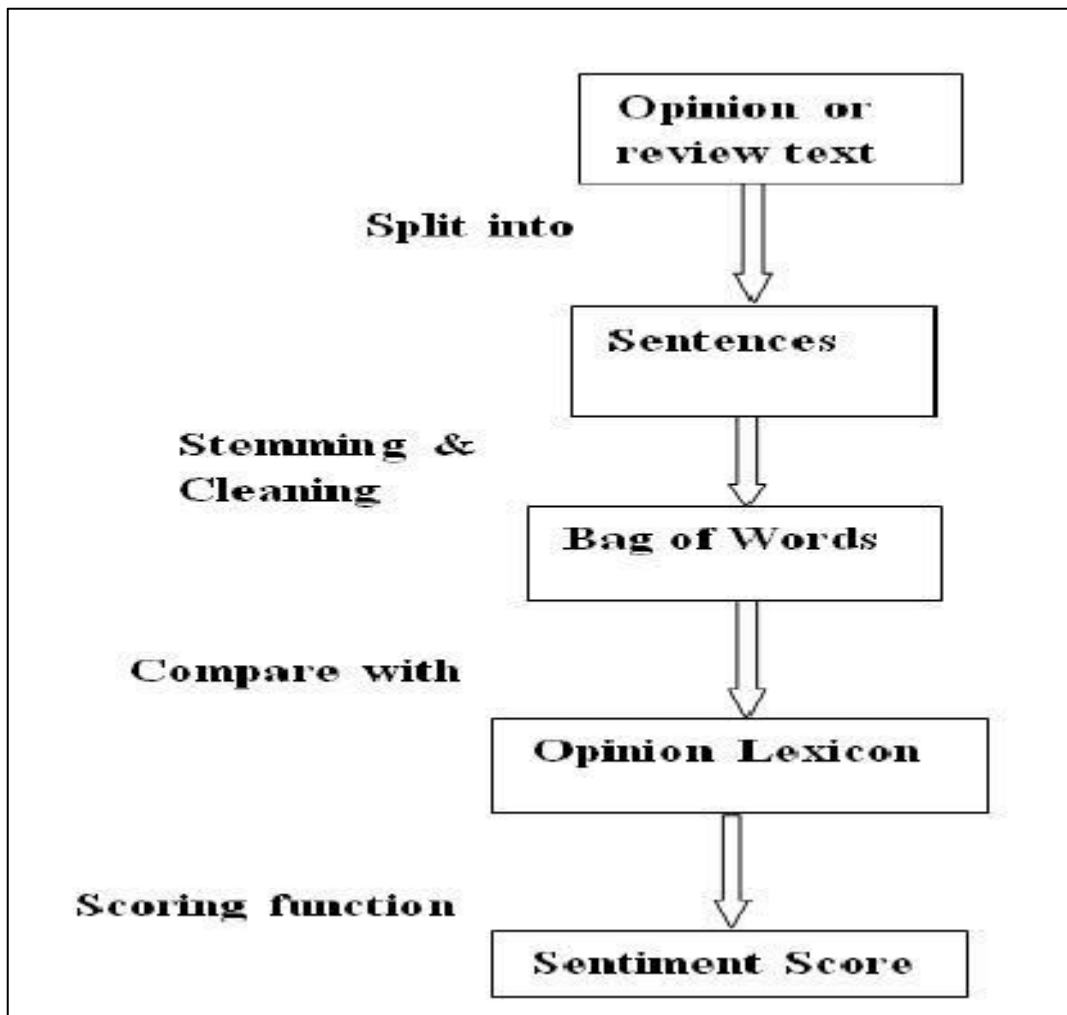


Figure 1. Lexicon based Sentiment Analysis approach

- *Automatic Approaches*

This type of method does not rely on human-crafted rules but relies on machine learning techniques. A task is modelled as a classification problem, whereby a classifier is supplied a text and returns a category, such as Neutral, Negative or Positive. The classification step involves a statistical model like Naïve Bayes, Logistic Regression, Support Vector Machines (SVM) or Neural Networks to predict the category of the text provided.

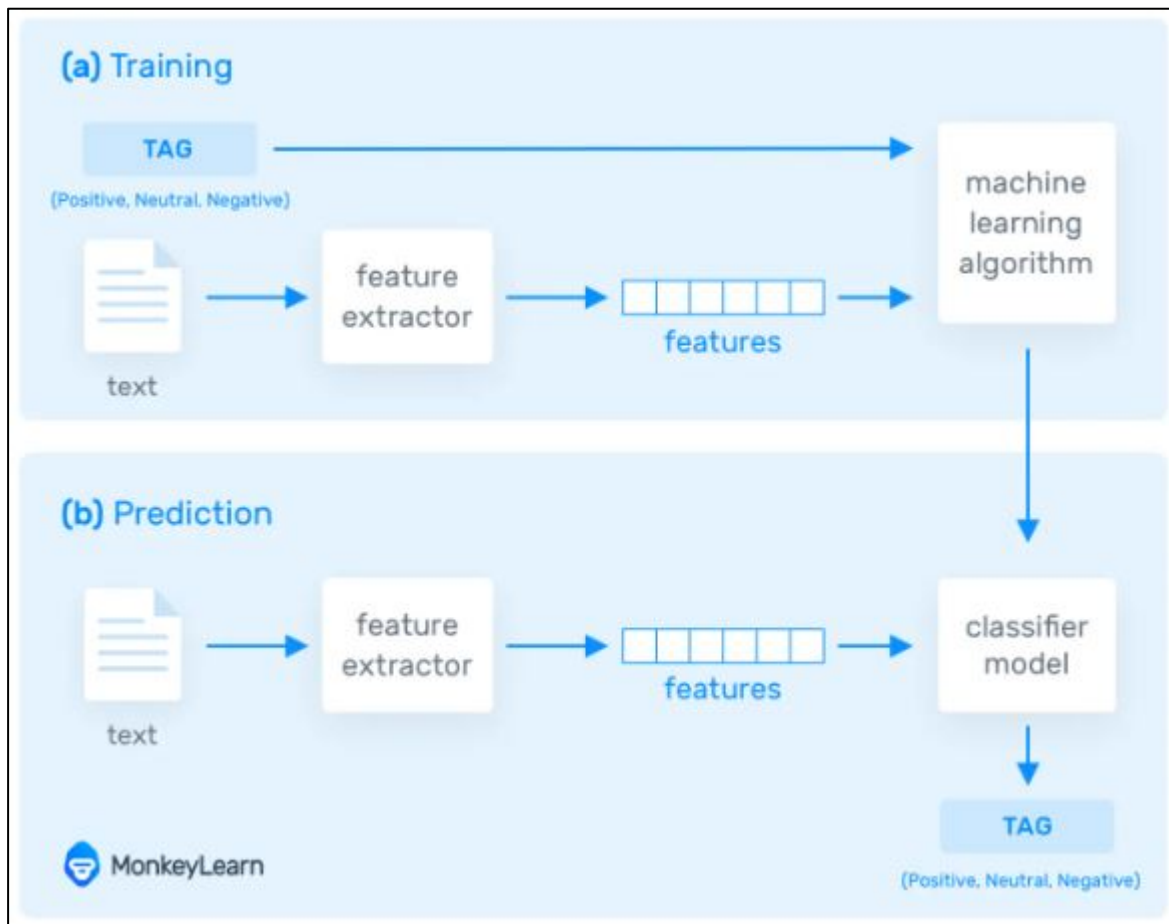


Figure 2. Machine learning classifier implementation

- *Hybrid Approaches*

This type of system combines the best elements of rule-based and automatic approach into single system resulting in more accurate results.

## 2 METHODOLOGY

### 2.1 Introduction to CRISP-DM

The CRISP-DM methodology will be used for the purpose of this project.

The CRISP-DM methodology offers a structured approach to designing a data mining project and is a well-established and robust methodology. CRISP-DM stands for Cross-Industry Process for Data Mining.(Europe)

The model shown below (Figure 3.) is an idealized cycle of events and several of the tasks can be carried out in a different order and it will every so often be required to backtrack to earlier tasks and reiterate certain actions. The model does not try to secure all potential routes through the data mining process.(Europe)

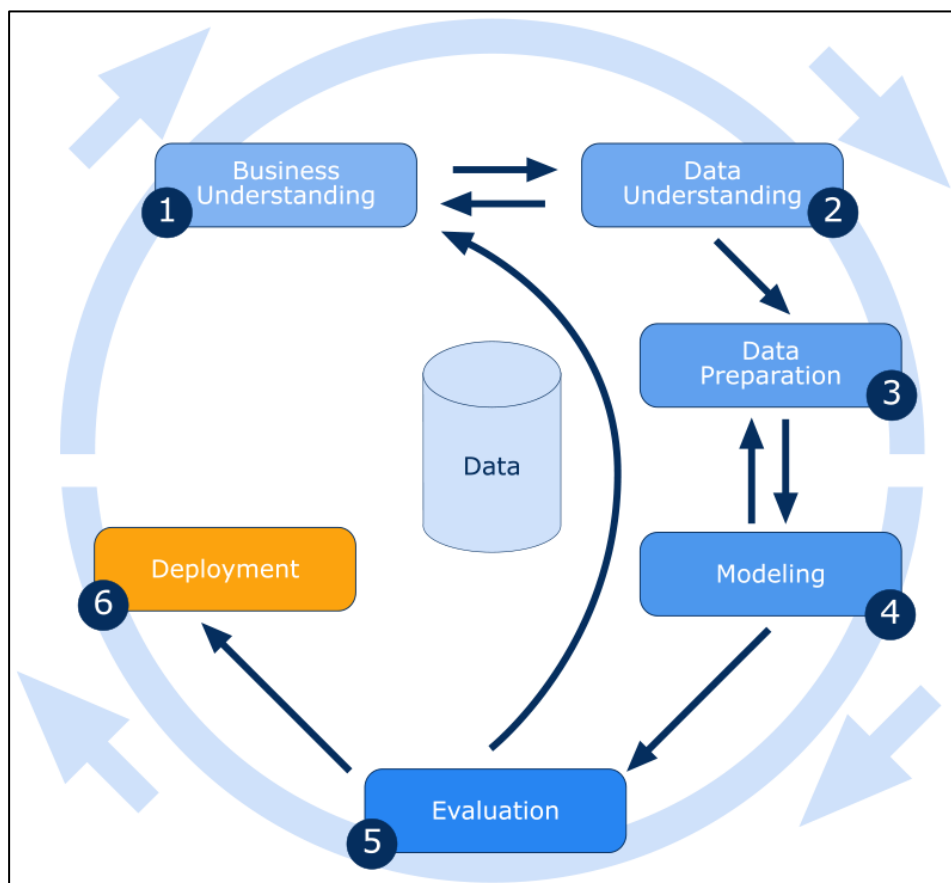


Figure 3. CRISP-DM Methodology

There are mainly 6 steps involved in Data Mining, which are as follows(Europe) :

#### Business Understanding

This step involves understanding the business objectives and understand what exactly the business requires. We assess the current situation by finding the resources, assumptions, limitations and other crucial factors to be considered. A good data mining strategy should be created to accomplish the business objectives.

#### Data Understanding

This phase begins with the preliminary data collection, which is acquired from open data sources, to get familiar with the data. The gross properties of the obtained data need to be checked carefully and reported. The data needs to be then studied and the quality of the data needs to be assessed. This step helps in confronting data mining questions and some other questions such as missing values in the data set.

#### Data Preparation

The data acquired from the data sources needs to be prepared, which includes selection, cleaning, constructing and formatting in the desired format.

#### Data Modelling

In this phase, firstly the modelling technique needs to be selected for the prepared data set and then the test scenario must be produced to confirm the quality and validity of the model. Example of modelling techniques are Decision tree, neural network generation etc.

## Evaluation

In this phase, the model findings must be evaluated in the framework of the business objectives which was set in the first phase. New patterns might be discovered in the model results which may prompt raising new business requirements. Gaining business knowledge is an iterative process.

## Deployment

The knowledge or information gained through the mining process, needs to be presented to the stakeholders so that they can make use of it. The deployment phase from business perspective could be just creating a report or complex reiterative data mining process across the business. In this phase, plans for deployment, maintenance and monitoring must be planned for implementation and future support.

## 2.2 COVID-19 tweets sentiment analysis

### 2.2.1 Business Understanding

Since WHO announced the coronavirus (Covid-19) pandemic as a Public Health Emergency of International Concern (PHEIC) on 30th January 2020, there has been varied widespread sentiment across the whole globe (*WHO Timeline - COVID-19, 2020*). There has been a massive impact on global economies due to the spread of COVID-19 resulting in growing number of business closures and loss of employment. According to International Monetary Fund (IMF), an estimated 10.4% proportion of people in the United States are out of work and millions of workers all over the world have either been laid off or have been put on government-funded job retention packages as some of the industries such as tourism, aviation or hospitality came to an abrupt halt due to the unfortunate yet essential lockdowns.(Lora Jones, 2020) The stock market has not been spared either where FTSE, DOW and NIKKEI all took a hit from the beginning of March 2020 with some positive recovery recently due to the some positive news in some parts of the world and gradual re-opening of businesses.(Lora Jones, 2020) The retail industry also took a hit as the shoppers preferred to stay at home and shop online(Lora Jones, 2020) where company like Amazon has benefitted due to its cloud computing capabilities.(Kari Paul, 2020)

The Irish economy has also not been spared of the negative impact of the COVID-19 pandemic. There was a decline in employment figures recorded approximately at 320,000 as of 27<sup>th</sup> March 2020. The Central Bank of Ireland expects an 8.3% decline in the real GDP due to the impact of interruption to businesses.(Kelly C De Bruin, 2020) The main events which happened during this period are: first case of coronavirus in Ireland was confirmed on 29<sup>th</sup> February 2020, 12<sup>th</sup> March 2020 - educational institutions close – partial lockdown, 27<sup>th</sup> March 2020 - Nationwide lockdown for 2 weeks, 18<sup>th</sup> May 2020 – Phase 1 and 8<sup>th</sup> June 2020 saw the ease in COVID-19 restrictions in Ireland.(Times, 2020; TeleTradar, 2020; Gov.ie, 2020)

Because of the ongoing public health and economic crisis, the general public in Ireland has been facing a lot of problems and a lot of emotions either positive, negative or neutral can be expected. Therefore, this brings us to the questions which we want to answer through this assignment:

1. What is the overall mood of the people in the 11 most populous counties in Ireland during the COVID-19 pandemic?
2. Can we find any fluctuation in the sentiment of the people in this period of the COVID-19 pandemic in Ireland?
3. Can we identify the sentiment of the people, which might help businesses to strategize accordingly?

### 2.2.2 Data Understanding

For the purpose of the project, tweets consisting of COVID related hashtags were sourced from twitter using a python script with a package known as GetOldTweets3 (See Appendix 1). The dataset consisted of tweets consisting of atleast one of the following words mentioned: 'Covid19', 'covid', 'corona', 'stayathome', 'lockdown' or 'coronavirus'. The time period the tweets are related to are from '01-03-2020' to '15-07-2020', which corresponds to the time from when Ireland detected its first positive COVID-19 case to the gradual relaxation of the countrywide restrictions. The tweets have been compiled from the 11 most populous counties of Ireland namely; Co. Dublin, Co. Cork, Co. Limerick, Co. Wexford, Co. Wicklow, Co. Donegal, Co. Galway, Co. Kerry, Co. Kildare, Co. Meath and Co. Tipperary, which would provide an overall picture regarding the sentiment of the nation in the times of COVID-19 pandemic. The scraped data was then written in a dataframe which was then compiled

into different .csv files county-wise and saved on the local hard drive. The names of the columns in the datasets were as follows: 'Datetime', 'Text', 'Hashtag' and 'Retweets', but the row number was different for each .csv file.

Issues/Limitations with the dataset:

- The different county-wise .csv files needed to be merged into one large .csv file for the purpose of analysis.
- The variable or column name 'Datetime' consisted of both date and time and would need to be separated into two separate columns.
- There was lack of a 'Location' variable or column, even though we had scraped county-wise data using python script.
- The number of tweets were not equal for each of the counties.

	A	B	C	D
1	Datetime	Text	Hashtag	Retweets
2	2020-07-14 21:31:06+00:00	Today training @earlyyears worke	#COVID19	2
3	2020-07-14 20:56:10+00:00	This actually did shock me #Corona	#CoronaVirus #Covid19 #Coro	0
4	2020-07-14 19:59:59+00:00	Rather than through a `eureka` mo	#COVID19 #clinicaltrial #teck	0
5	2020-07-14 19:47:09+00:00	#CoronaVirus #Covid19	#CoronaVirus #Covid19	0
6	2020-07-14 19:44:47+00:00	No need. I had followed other clues, and knew the covid19 was		0
7	2020-07-14 19:36:17+00:00	Looking for examples of good #CO	#COVID19 #Limerick	3
8	2020-07-14 19:26:03+00:00	Just caught a glimpse of myself w	#facemask #COVID19	0
9	2020-07-14 19:21:17+00:00	Assisting at this evening's Irish Blc	#limerickcivildefence #ourvol	0
10	2020-07-14 18:59:59+00:00	The sheer volume of #clinicaltrials	#clinicaltrials #COVID19 #tec	0
11	2020-07-14 18:56:25+00:00	Denis Reen is a dentist, a man of s	#covid19	0
12	2020-07-14 18:50:00+00:00	In case you missed it: As scheduled	#StaySafe #COVID19 #Midwe	4
13	2020-07-14 18:19:49+00:00	The Govt has put in place a signific	#COVID19	2
14	2020-07-14 17:50:46+00:00	32 new cases of #COVID19 today.	#COVID19	0
15	2020-07-14 14:29:16+00:00	COVID19 has had a big impact on	#mentalhealth	1
16	2020-07-14 13:00:10+00:00	Clinical trial continuity is a challeng	#COVID19 #coronavirus #COV	0
17	2020-07-14 12:21:00+00:00	The lockdown `really showed how	#MIDLimerick #COVID19	0
18	2020-07-14 12:04:42+00:00	What a waste of money by those donating. Keep your money sa		0
19	2020-07-14 08:59:59+00:00	With heightened awareness of clir	#COVID19 #teckrotheanswer	0
20	2020-07-14 08:45:33+00:00	Great article by Limerick based do	#stayhome #stayhome #stay	0

Figure 4. Snapshot of data from the Limerick.csv file

### 2.2.3 Data Preparation

In order to progress with our analysis and improve the data quality, the data needed to be transformed and then cleaned.

An additional variable/column named, 'Location' was added manually to each of the individual .csv files using MS Excel (Figure5.).

	A	B	C	D	E
1	Datetime	Text	Hashtag	Retweets	Location
2	2020-07-14 23:57:56+00:00	Champion level safety! @LFC	#YNWA #COVID19	0	Dublin
3	2020-07-14 23:51:30+00:00	The latest World News! https://	#eu #covid19	0	Dublin
4	2020-07-14 23:46:21+00:00	More likely god is angry with Trump, that's why		0	Dublin
5	2020-07-14 23:36:41+00:00	If the media (or other reliable	#covid19	0	Dublin
6	2020-07-14 23:21:43+00:00	#TruthBeTold ABOUT THE #CC	#TruthBeTold #CC	0	Dublin
7	2020-07-14 23:18:40+00:00	I just came across on Instagram many Iranian st		0	Dublin
8	2020-07-14 23:12:14+00:00	someone cut open covid19 and make sure it isn't		1	Dublin
9	2020-07-14 22:47:05+00:00	No country so far has managed to reopen school		0	Dublin
10	2020-07-14 22:31:16+00:00	Delighted to be interviewed on	#covid19	1	Dublin
11	2020-07-14 22:22:19+00:00	Couldn't agree with you more. People who only		0	Dublin
12	2020-07-14 22:10:48+00:00	Watching the press conference	#NewZealand #C	0	Dublin
13	2020-07-14 22:09:37+00:00	The Late Late Toy Show is going	#COVID19	0	Dublin
14	2020-07-14 22:09:33+00:00	Collective Teacher Efficacy - r	#CollectiveTeache	1	Dublin
15	2020-07-14 22:06:58+00:00	https://m.facebook.com/story	#Venezuela #mac	0	Dublin
16	2020-07-14 22:06:42+00:00	There are very few remaining parents left of the		9	Dublin
17	2020-07-14 22:02:21+00:00	Paying attention to business c	#QIreland	0	Dublin
18	2020-07-14 22:00:01+00:00	#WATCH Lack of guidelines fo	#WATCH #VMNew	0	Dublin
19	2020-07-14 21:58:55+00:00	#COVID19 has made it harder	#COVID19 #Parkir	1	Dublin
20	2020-07-14 21:58:12+00:00	Now #France2tv is rebroadcast	#France2tv #conc	1	Dublin

Figure 5. New Column 'Location' created in the .csv file

A R-script (See Appendix 2) was written in order to merge the 11 different .csv files into 1 large .csv file.

The merged .csv file consisted of 5 columns initially and 14,974 rows. Column names were as follows: 'Datetime', 'Text', 'Hashtag' and 'Retweets' and 'Location'. There were no missing values in the dataset. Next, for the ease of analysis, the 'Datetime' column was separated into 2 different columns namely, 'Date' and 'Time' using the Text to Columns function in MS Excel with Space being the delimiter. The column 'Retweets' was removed from the dataset as it was not going to be used in the analysis.

	A	B	C	D	E
1	Date	Time	Text	Hashtag	Location
2	01/03/2020	23:56:27+00:00	Day 0/14 of WFH #COVID19 I am ser	#COVID19	Dublin
3	01/03/2020	23:38:48+00:00	Yes! @1GaryGannon Respect to hea	#COVID19	Dublin
4	01/03/2020	23:37:27+00:00	Right, working tomorrow so off to sl	#PulpFicton #COVID19	Dublin
5	01/03/2020	22:51:21+00:00	But But But #Covid19 ! Have you nev	#Covid19	Dublin
6	01/03/2020	21:38:03+00:00	So the department of health isnít na	#covid19 #Coronavirius	Dublin
7	01/03/2020	20:43:48+00:00	As ever, a thread full of good sense	#washyourhands #wash	Dublin
8	01/03/2020	20:33:24+00:00	At least use the right hashtag #Coro	#Coronavirus #Covid19	Dublin
9	01/03/2020	18:38:57+00:00	I understand this is a scary time for	#COVID19	Dublin
10	01/03/2020	18:32:03+00:00	Does #COVID19 really want to take c	#COVID19 #CoronaVirus	Dublin
11	01/03/2020	18:15:39+00:00	What countries have been affected	#coronavirus #COVID19	Dublin
12	01/03/2020	18:14:23+00:00	We are conscious of our responsibili	#COVID19	Dublin
13	01/03/2020	17:43:45+00:00	#EU News: #Dublin secondary #scho	#EU #Dublin #school #st	Dublin
14	01/03/2020	17:34:06+00:00	A secondary school in "the east of Ir	#coronavirus #covid19	Dublin
15	01/03/2020	16:44:05+00:00	Amongst all the #covid19 hysteria ar	#covid19 #covid19	Dublin
16	01/03/2020	15:15:28+00:00	The great pasta famine is upon us..	#COVID19 #Coronavirusi	Dublin
17	01/03/2020	14:49:55+00:00	In #blanchardstown centre today anc	#blanchardstown #COVI	Dublin
18	01/03/2020	12:58:40+00:00	#COVID19 #Coronavid19 Important i	#COVID19 #Coronavid19	Dublin
19	01/03/2020	12:15:11+00:00	@RTERadio1 panellist this morning	#COVID19	Dublin
20	01/03/2020	11:41:00+00:00	It is so cold these days... Lots of viru	#trusTEA #flu #viruses #	Dublin
21	01/03/2020	11:36:55+00:00	Excellent thread with hearty referen	#SARSCoV2 #HCoV19 #C	Dublin
22	01/03/2020	11:26:56+00:00	@WeekendOnOneRTE Young people	#COVID19	Dublin
23	01/03/2020	11:21:45+00:00	My own mother wonít even tell me	#COVID19	Dublin
24	01/03/2020	11:17:12+00:00	First case of COVID-19 identified in f	#coronavirus #covid19 #	Dublin
25	01/03/2020	01:38:56+00:00	Is #COVID19 a signal from mother e	#COVID19	Dublin
26	01/03/2020	23:41:44+00:00	A million dead, worst case scenario.	#COVID19	Galway
27	01/03/2020	20:22:06+00:00	There's a difference between homes	#COVID19 #homeschool	Galway

Figure 6. Final transformed dataset

The final merged and transformed dataset consisted of 5 columns and 14,974 rows.

Next, we performed data cleaning, created a 'sentiment' column with polarity score and labelled our tweets based on the polarity score in an another column 'mood'. All this was performed by writing an R-script and running it (Appendix 3).

#### a. Data Cleaning

The required libraries were installed and loaded in R. The final dataset .csv was read and a dataframe was created. The 'Text' column consisting of the tweets needed to be cleaned which was done using the gsub() function in R. The following steps were taken to clean the tweets:

- Converted all text into lower case and replaced the http links with "".
- Replaced punctuations from the text to "".

- Replaced alphanumeric words and digits from the text to "".
- Replaced string 'rt' and '@' with "".
- Replaced the most common words like 'covid', 'coronavirus' etc with "".

Post cleaning the tweets, there were few rows of data which were left blank. These were first converted to 'NA' and then omitted from the dataset.

b. Column 'sentiment'

Next, sentiment analysis was performed by loading and using the library 'sentimentr' in R. Each tweet was given a sentiment score and a 'sentiment' column/variable was created.

c. Column 'mood'

Next, the tweets needed to be labelled based on their sentiment scores. This was done again using R. A column/variable 'mood' was created where tweets with sentiment score above 0 were labelled 'Positive', sentiment score below 0 were labelled as 'Negative' and rest were labelled as 'Neutral'.

Finally, the labelled dataframe was written into a .csv file, which can be used for modelling. The total no. of columns in the labelled dataset was 8 namely, 'ID', 'Date', 'Time', 'Text', 'Hashtag', 'Location', 'sentiment', 'mood' and the total no. of rows was 14,921.

A	B	C	D	E	F	G	H
ID	Date	Time	Text	Hashtag	Location	sentiment	mood
1	01/03/2020	23:56:27+00:	day of wfh i am seriously concerned a	#COVID19	Dublin	-0.0785584	Negative
2	01/03/2020	23:38:48+00:	yes garygannon respect to healthcare p	#COVID19	Dublin	0.38031942	Positive
3	01/03/2020	23:37:27+00:	right working tomorrow so off to sleep	#PulpFictio	Dublin	0.22188008	Positive
4	01/03/2020	22:51:21+00:	but but but have you never watched th	#Covid19	Dublin	0.31622777	Positive
5	01/03/2020	21:38:03+00:	so the depament of health isnt naming	#covid19 #Cc	Dublin	0	Neutral
6	01/03/2020	20:43:48+00:	as ever a thread full of good sense fron	#washyourh	Dublin	0.22613351	Positive
7	01/03/2020	20:33:24+00:	at least use the right hashtag or	#Coronavirus	Dublin	0.06047432	Positive
8	01/03/2020	18:38:57+00:	i understand this is a scary time for ma	#COVID19	Dublin	0.96307883	Positive
9	01/03/2020	18:32:03+00:	does really want to take on the nohsid	#COVID19 #C	Dublin	0	Neutral
10	01/03/2020	18:15:39+00:	what countries have been affected by t	#coronavirus	Dublin	0.02	Positive
11	01/03/2020	18:14:23+00:	we are conscious of our responsibility t	#COVID19	Dublin	0.55339859	Positive
12	01/03/2020	17:43:45+00:	eu news dublin secondary school to clos	#EU #Dublin	Dublin	0	Neutral
13	01/03/2020	17:34:06+00:	a secondary school in the east of has b	#coronavirus	Dublin	-0.1405564	Negative
14	01/03/2020	16:44:05+00:	amongst all the hysteria and run on ha	#covid19 #cc	Dublin	-0.1356801	Negative
15	01/03/2020	15:15:28+00:	the great pasta famine is upon us no p	#COVID19 #C	Dublin	0	Neutral
16	01/03/2020	14:49:55+00:	in blanchardstown centre today and ive	#blanchardst	Dublin	0	Neutral
17	01/03/2020	12:58:40+00:	vid impoant information from the hse	#COVID19 #C	Dublin	0.08944272	Positive
18	01/03/2020	12:15:11+00:	eradio panellist this morning differenti	#COVID19	Dublin	-0.2150349	Negative
19	01/03/2020	11:41:00+00:	it is so cold these days lots of es in the	#trusTEA #fl	Dublin	-0.2086997	Negative
20	01/03/2020	11:36:55+00:	excellent thread with heay references a	#SARSCoV2	Dublin	0.31622777	Positive

Figure 7. Final cleaned and labelled dataset

The final labelled dataset was then balanced in RapidMiner for the purpose of Modelling in RapidMiner. The process can be seen below which was used to sample, append and then store the balanced dataset in a .csv file namely, 'labelled one final.csv'. The sample size of each category, 'Positive', 'Negative' and 'Neutral' was 2,174.

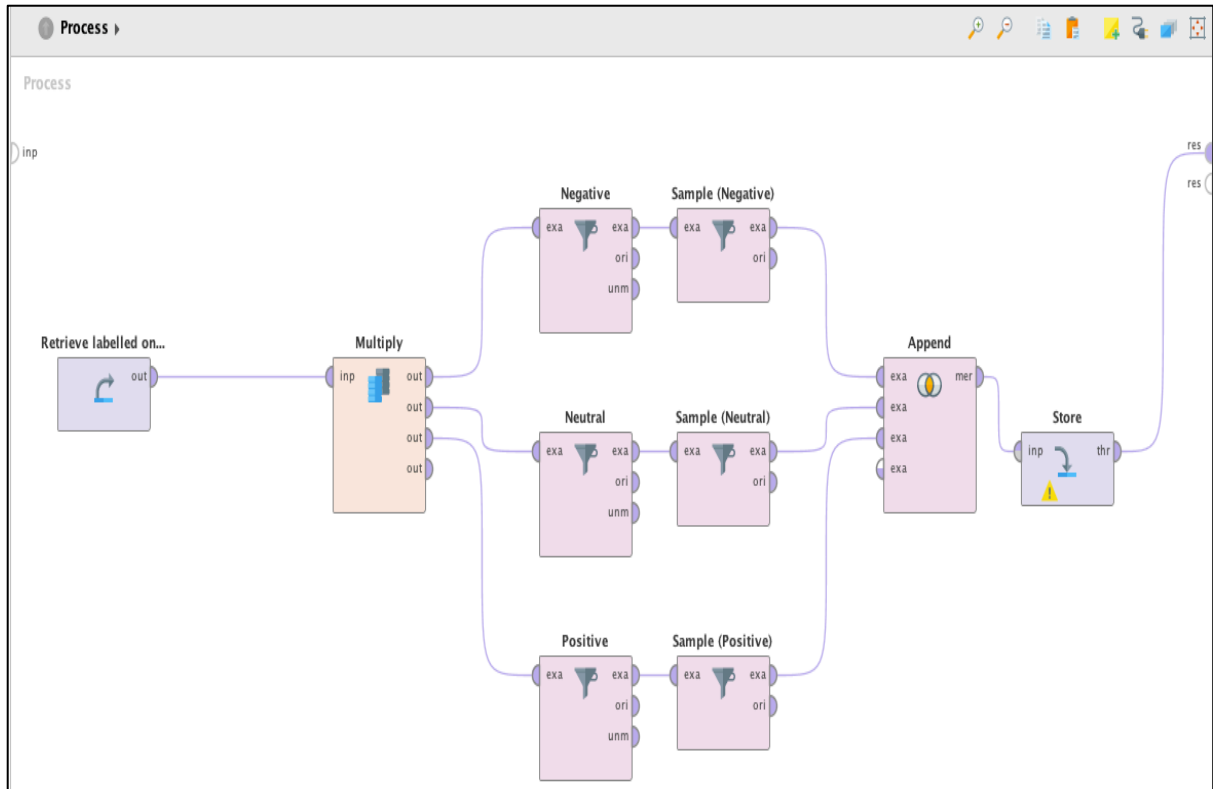


Figure 8. Process in RapidMiner for sampling and balancing the dataset

## Text Vectorization

Text Vectorization is a process of converting text into numerical features. Text vectorization is done automatically in Rapidminer during the modelling process (Figure 9). We will now be discussing some of the terminology used in Vectorization, which are as follows: (Hoare, 2020; Chen, 2020; Raschka, 2014)

- Term Document Matrix (TDM)

It is a document vector where each document is an example and each token is an attribute. Either term counts or term frequencies or TDF scores can occur in the cells.

- Stopwords

Stopword filtering is employed to remove or filter out the most common words in a language. For example, 'a', 'the', 'that', 'and' etc. This process significantly reduces the size of TDM.

- Stemming

It is the process of trimming each token to its most essential minimum.



## Cross-validation

Cross-validation is a technique in which we train our model using the subset of the dataset and then calculate using the complementary subset of the dataset. There are 3 steps present in cross-validation, which are as follows: (Sharma, 2020)

1. Keep some proportion of the sample dataset.
2. Training the model on the rest of the dataset.
3. Using the reserve proportion of the dataset to test the model.

A popular and efficient method of cross-validation is the K-Fold Cross-validation which we will be discussing in brief.

In K-Fold Cross Validation, the dataset is split into k number of folds or subsets and then training is performed on all the folds but leave one (k-1) subset for the evaluation of the trained model. We iterate k times with a dissimilar subset kept for testing purpose each time. As a general rule and empirical evidence,  $K = 5$  or  $10$  is generally desired but it is not set in stone and any value of  $K$  can be used.(Gupta, 2017)

The main advantage of this method is that it runs  $K$  times faster because K-Fold cross validation repeats the train/test split  $K$ -times. It presents a more accurate estimate and also every observation in the data is used for both training and testing making it a more efficient method.(Sharma, 2020)

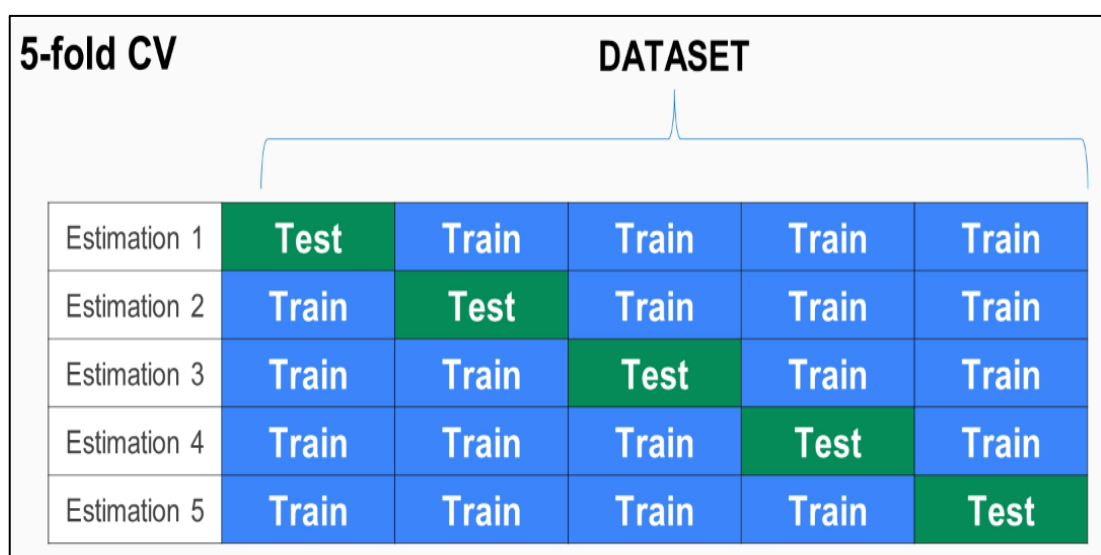
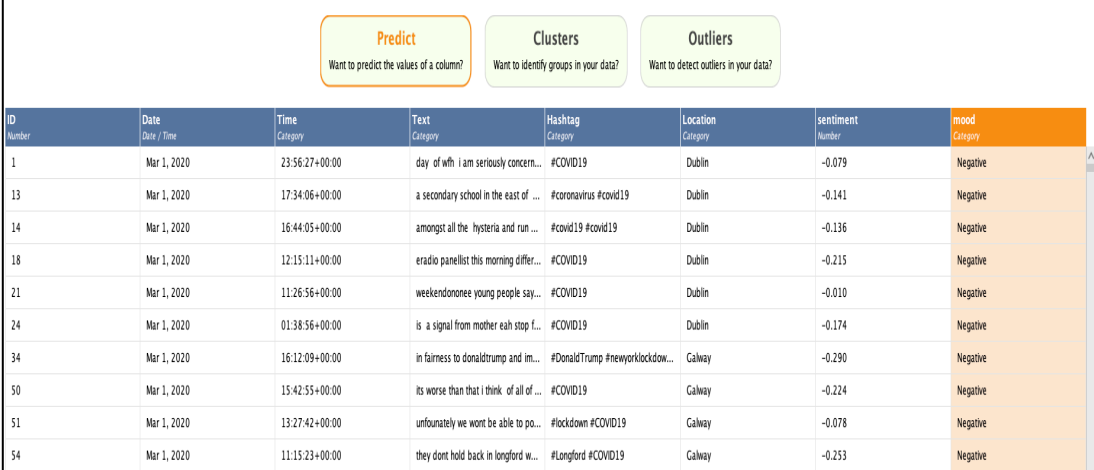


Figure 10. Example of 5-fold cross validation

## 2.2.4 Data Modelling and Evaluation

After the data preparation phase, the balanced dataset namely, 'Balanced covid data' was loaded into RapidMiner. The Auto Model feature of RapidMiner was used to observe which model would yield the best results for our dataset. The Auto Model involves different steps which involves the following:

- Loading/Importing the balanced dataset
- Select the task 'Predict' and choosing the column 'mood', which has 3 categories, Positive, Negative and Neutral.

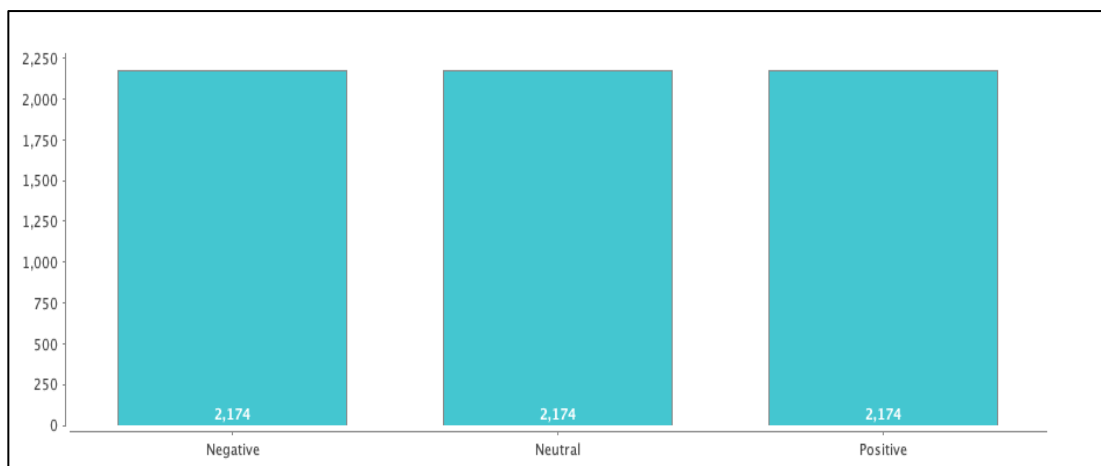


The screenshot shows the RapidMiner Auto Model interface. At the top, three buttons are visible: 'Predict' (highlighted in orange), 'Clusters', and 'Outliers'. Below these buttons is a table with the following columns: ID Number, Date Date / Time, Time Category, Text Category, Hashtag Category, Location Category, sentiment Number, and mood Category. The 'mood' column is highlighted in orange, indicating it is the target variable for prediction. The table contains 10 rows of data, all with a 'Negative' mood category.

ID Number	Date Date / Time	Time Category	Text Category	Hashtag Category	Location Category	sentiment Number	mood Category
1	Mar 1, 2020	23:56:27+00:00	day of wfh i am seriously concern...	#COVID19	Dublin	-0.079	Negative
13	Mar 1, 2020	17:34:06+00:00	a secondary school in the east of ...	#coronavirus #covid19	Dublin	-0.141	Negative
14	Mar 1, 2020	16:44:05+00:00	amongst all the hysteria and run ...	#covid19 #covid19	Dublin	-0.136	Negative
18	Mar 1, 2020	12:15:11+00:00	eradio panellist this morning differ...	#COVID19	Dublin	-0.215	Negative
21	Mar 1, 2020	11:26:56+00:00	weekendononee young people say...	#COVID19	Dublin	-0.010	Negative
24	Mar 1, 2020	01:38:56+00:00	is a signal from mother eah stop f...	#COVID19	Dublin	-0.174	Negative
34	Mar 1, 2020	16:12:09+00:00	in fairness to donaldtrump and im...	#DonaldTrump #newyorklockdown...	Galway	-0.290	Negative
50	Mar 1, 2020	15:42:55+00:00	its worse than that i think of all of ...	#COVID19	Galway	-0.224	Negative
51	Mar 1, 2020	13:27:42+00:00	unfounately we wont be able to po...	#lockdown #COVID19	Galway	-0.078	Negative
54	Mar 1, 2020	11:15:23+00:00	they dont hold back in longford w...	#Longford #COVID19	Galway	-0.253	Negative

Figure 11. Selecting the task and column to predict values

- Selecting the significant input columns. The columns 'Text', 'sentiment', 'Date', 'Hashtag' and 'Location' were selected as input variables.



Selected	Status ↑	Quality	Name	Correlation	ID-ness	Stability	Missing	Text-ness
<input type="checkbox"/>	<span style="color: red;">●</span>		ID	0.79%	100.00%	0.02%	0.00%	0.00%
<input type="checkbox"/>	<span style="color: red;">●</span>		Time	62.04%	94.96%	0.06%	0.00%	37.87%
<input checked="" type="checkbox"/>	<span style="color: yellow;">●</span>		Text	66.11%	97.90%	0.65%	0.49%	98.78%
<input checked="" type="checkbox"/>	<span style="color: yellow;">●</span>		sentiment	50.02%	?	33.33%	0.00%	0.00%
<input checked="" type="checkbox"/>	<span style="color: green;">●</span>		Date	0.80%	?	3.83%	0.00%	0.00%
<input checked="" type="checkbox"/>	<span style="color: green;">●</span>		Hashtag	16.82%	49.89%	23.34%	9.67%	54.94%
<input checked="" type="checkbox"/>	<span style="color: green;">●</span>		Location	0.08%	0.17%	70.25%	0.00%	2.77%

Figure 12. Selecting the input variables

- Selecting Model types. In the 'Model Types' phase, we selected the models recommended by RapidMiner. After this, the modelling process was started.

## Models

Naive Bayes

Generalized Linear Model

Use Regularization     Calculate p-Values

Logistic Regression

Fast Large Margin

Automatically Optimize

Deep Learning

Decision Tree

Automatically Optimize    Maximal Depth:

Random Forest

Automatically Optimize    Number of Trees:     Maximal Depth:

Gradient Boosted Trees

Automatically Optimize    Number of Trees:     Maximal Depth:     Learning Rate:

Support Vector Machine

Automatically Optimize

Figure 13. Selecting the recommended models in RapidMiner

- Results

The results were obtained after all the models were run. We can see the accuracy and classification error of each type of model in Figure. 12 and 13. The results have been summarised in a table for the ease of observation in Figure. 14.

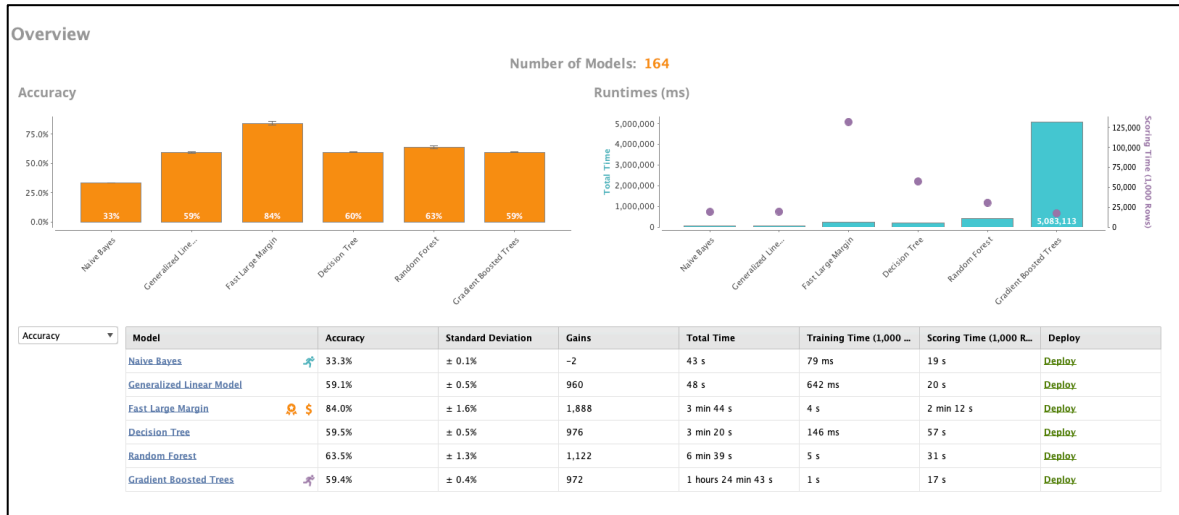


Figure 14. Accuracy for each type of model

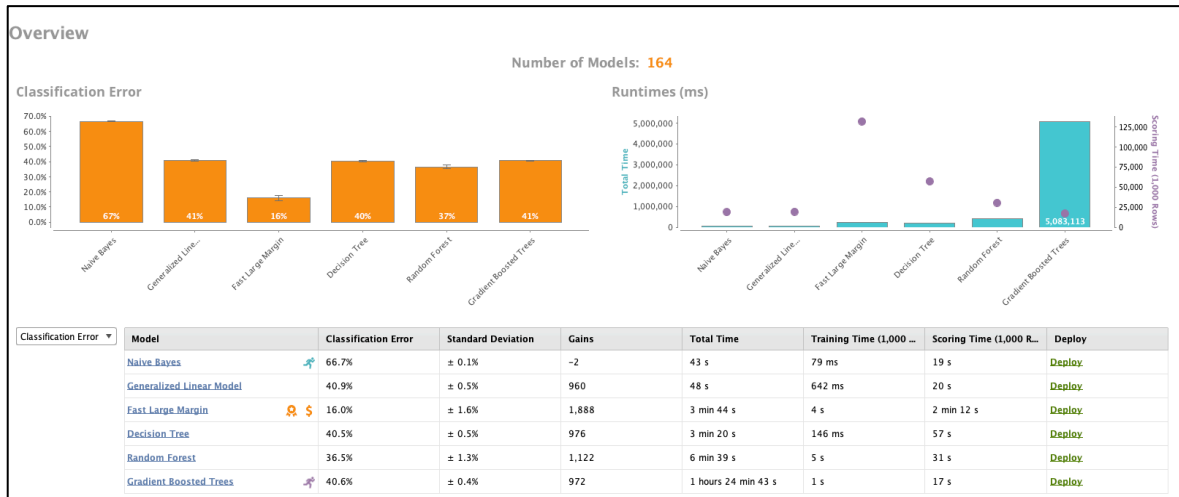


Figure 15. Classification error for each type of model

<b>Model</b>	<b>Classification Error</b>	<b>Accuracy</b>	<b>Standard Deviation (+/-)</b>	<b>Gains</b>	<b>Total Time</b>	<b>Training Time (1000 rows)</b>	<b>Scoring Time</b>
Naïve Bayes	66.7%	33.3%	0.1%	-2	00:00:45	00:00:00.117	00:00:18
Generalized Linear Model	40.9%	59.1%	0.6%	960	00:00:46	00:00:00.584	00:00:17
Fast Large Margin	16.3%	83.7%	0.7%	1878	00:01:45	00:00:00.659	00:00:37
Decision Tree	40.5%	59.5%	0.5%	976	00:02:16	00:00:00.331	00:00:45
Random Forest	36.5%	63.5%	1.3%	1122	00:12:32	00:00:09.000	00:01:13
Gradient Boosted Trees	40.6%	59.4%	0.4%	972	01:49:05	00:00:02.000	00:00:38

Figure 16. Table of results of the Rapidminer automodel

In the Figure 14 above, we can see and discuss the results of our Auto-model and select the best modelling technique.

- We can observe that the model 'Naïve Bayes' took the least time to run i.e. 45 seconds and the model 'Gradient Boosted Trees' took the most time to run i.e. 1 hr 49min and 5 seconds. The second fastest model to run was the 'Generalized Liner Model'.
- For Accuracy, 'Fast Large Margin' achieved the highest score of 83.7% when compared to the second best model i.e. 'Random Forest' which was 63.5%. Lowest accuracy was seen for the model 'Naïve Bayes' even though it took the least amount of time to run.
- For Classification error, it was lowest for the 'Fast Large Margins' and second lowest for the model 'Random Forest'.
- 'Naïve Bayes' model is the typical benchmark in text mining but in this particular task, 'Fast Large Margin' (SVM) seems to have outperformed.

The 'Fast Large Margin' Model yielded the most accurate model when compared to other modelling techniques. The Fast Large Margin is a fast learning method for large margin optimizations. The Fast Large Margin operator in Rapidminer applies a fast margin learner based on the linear support vector learning scheme advised by R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin. The results provided by this method is similar to classical SVM or logistic regression operation but this linear classifier is able to operate datasets consisting of millions of variables and examples.(Core, 2020)

We will be discussing few types of the algorithms used in Rapidminer.(Sidana, 2017)

## 1. Naïve Bayes

The algorithm is based on the Bayes' Theorem with assumption of independence among predictors. The algorithm assumes that the existence of a particular item in a class is unrelated to the presence of any other item or that all of these properties have independent involvement to the probability.

Naïve Bayes is known to be highly scalable and easy to build.

$$P(c|x) = \frac{P(x|c).P(c)}{P(x)}$$

where  $P(c|x)$  is Posterior probability,  $P(x|c)$  is Likelihood,  $P(c)$  is Class prior probability and  $P(x)$  is Predictor prior probability

## 2. Decision Trees

This algorithm builds classification or regression models in the form of a tree structure. It splits a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally created. A tree with decision nodes and leaf nodes is the final result. A decision node has 2 or more branches and a leaf node denotes a classification or decision. The chief decision node in a tree which corresponds to the best predictor is known as root node. It can process both categorical and numerical data.

## 3. Random Forest

It is an ensemble learning method for classification, regression and other tasks, that operate by creating an array of decision trees at training time and yielding the class that is the mode of the classes or mean prediction of the individual trees. They correct for the decision trees' problem of over fitting to their training set.

## 4. Generalized Linear Model

In Statistics, Generalized Linear Model (GLM) is a variable generalization of ordinary linear regression that permits for response variables that have error distributions other than normal.

GLM's make allowance for response variables that have arbitrary distributions (rather than simply normal distributions). They allow an arbitrary function (the link function) of the responsible variable to vary linearly with the predicted values (rather than if the response itself must vary linearly). The generalized linear model typically fits a model to the data to maximize the log-likelihood.(Wikipedia, 2020)

### 2.2.5 Deployment

As per the CRISP-DM process, the next step would be deployment. RapidMiner does have the functionality to deploy your model locally on your PC.

For the purpose of this project, we are going to deploy a PowerBI Dashboard to understand the sentiments around Covid-19 in the Republic of Ireland. The dashboard has been provided along with this report, so that the project becomes engaging to the user and they can themselves explore different insights on this subject matter.

Another benefit of using PowerBI is that it can be deployed on mobile devices.

## 3 RESULTS

We will now be discussing the final results gained by visualising the sentiment analysis performed. The Covid-19 tweet data is from a short period of time from 1<sup>st</sup> March 2020 to 15<sup>th</sup> July 2020. The visualisations have been prepared in Power BI except the wordcloud which has been created using R Studio.

In Figure 17, we can see percentage of Positive, Neutral and Negative tweets by location.

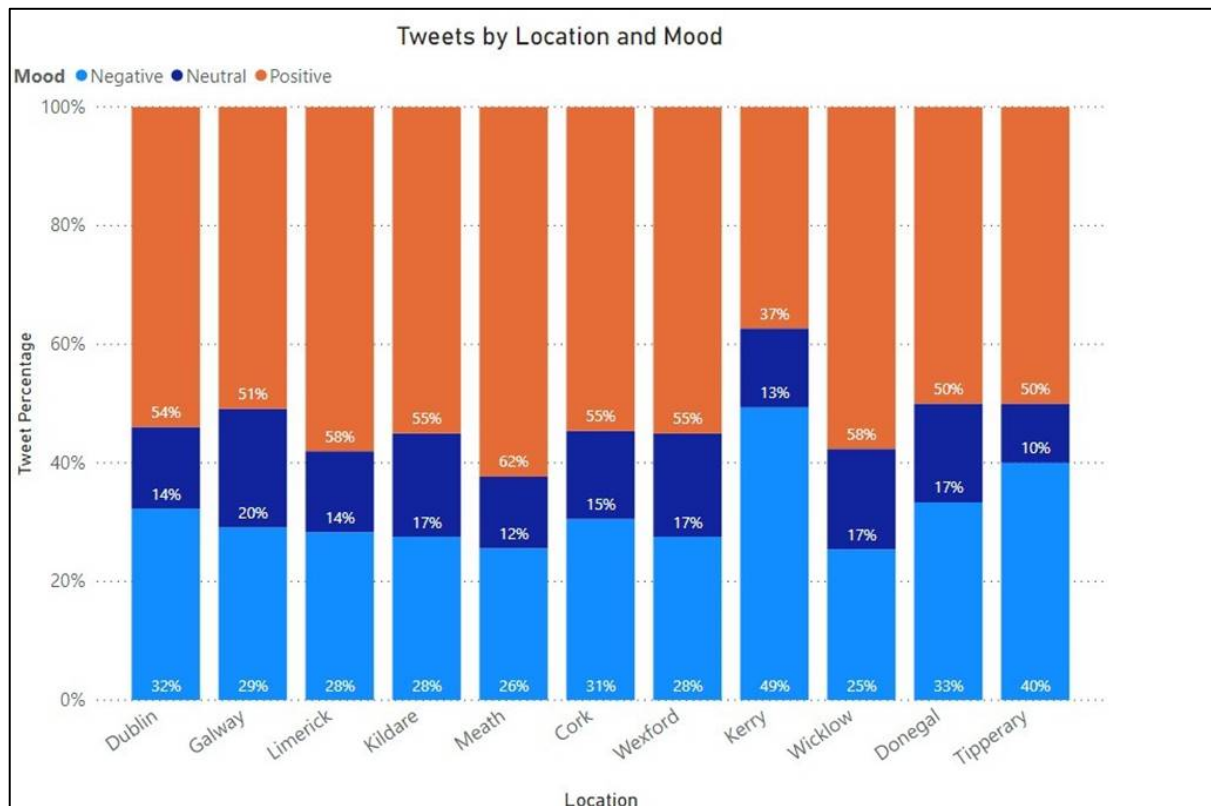


Figure 17. Tweets by Location and Mood

The highest percentage of positive tweets were seen in Co. Meath with 62% and second highest being Co. Limerick and Co. Wicklow being at 58% each. Co. Dublin was seen to have 54% of positive tweets.

Co. Kerry and Co. Donegal were seen to be with highest percentage of Negative tweets with 49% and 33% respectively. Co. Dublin was observed to have 32% of negative tweets.

At a quick glance, it can be inferred that across all the 11 most populous counties of Ireland, a positive sentiment could be seen regarding Covid-19.

In Figure 18, We can observe the no. of tweets which were positive, neutral and negative by date. The first tweet is from the 1<sup>st</sup> of March 2020 and the last tweet is from the 15<sup>th</sup> of July 2020. As we know, Ireland’s first case of coronavirus was reported in February 2020 from where a series of events took place and as a result a series of restrictions were put in place to curb the spread of the virus. We can see really high no. of negative tweets from the 1<sup>st</sup> of March 2020 but then it gradually dropped.

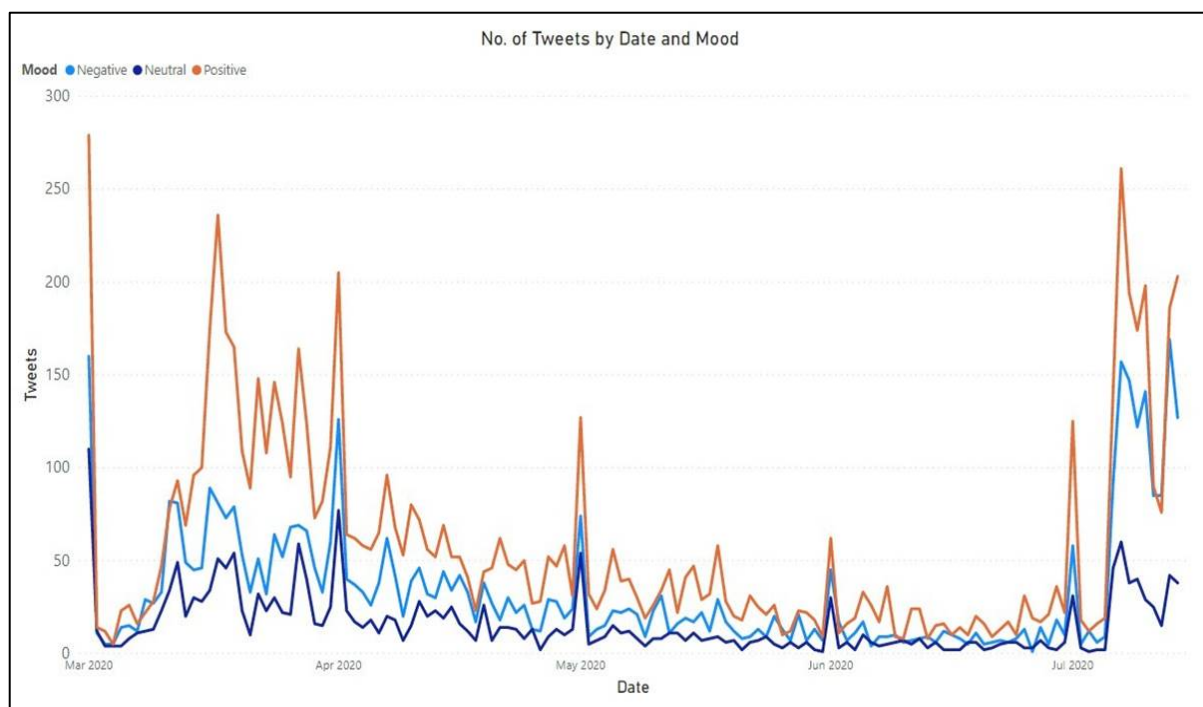


Figure 18. No. of tweets by Date and Mood

On 12<sup>th</sup> March 2020, a partial lockdown was announced with the closure of educational institutions i.e. schools, colleges and universities. In the figure, we can see a rise in negative tweets again from the 9<sup>th</sup> of March 2020 as the news of an imminent lockdown spread. On 27<sup>th</sup> March 2020, a countrywide 2 week lockdown was announced and as expected a rise in the no. of negative can be seen upto the 1<sup>st</sup> of April 2020.

On 18<sup>th</sup> May 2020, the country opened partially under Phase 1 and more restriction were eased with commencement of Phase 2. As from the figure, we can see a gradual decrease in numbers of positive and negative tweets and the reason could be attributed to the following reasons:

- Acceptance of the fact that Covid is here to stay and we have to work together to curb the spread.
- Gradual ease of restrictions from 18<sup>th</sup> May and onwards.
- People felt that there was an adequate response and support from the Irish government.

One point to note is the constant high number of positive tweets during this whole time period. This can be attributed to the fact the people of Ireland remained positive, hopeful and motivated even though the conditions in the country were deteriorating.

In figure 19, we can see the number of positive, neutral and negative tweets over the course 24 hours during the period from 1<sup>st</sup> March to 15<sup>th</sup> July 2020.

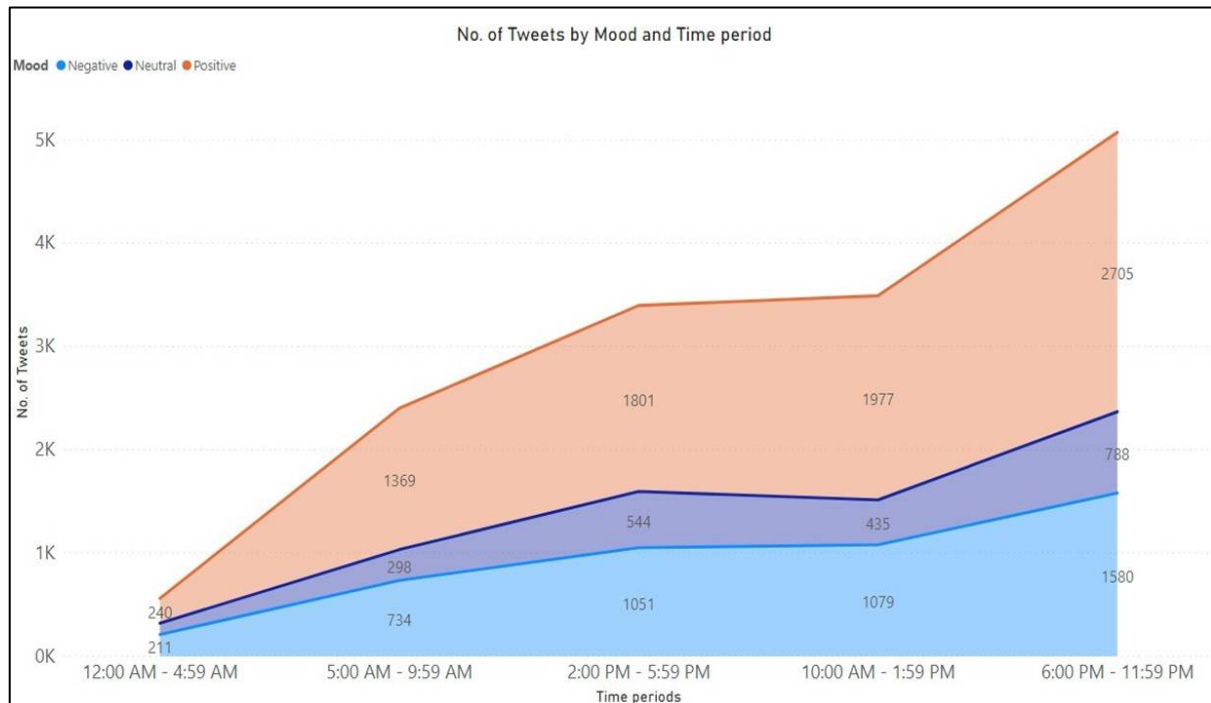


Figure 19. Tweets by Mood and Time

The visualisation above is very interesting to see as it provides an insight into the times people in 11 most populous counties in Ireland were tweeting about Covid-19. Here, we can observe that maximum number of tweets be it be positive, neutral and negative are tweeted in the hours between 6pm - 12pm. The next highest number of tweets were interestingly done during the hours 10am – 2pm and the reason could be that people getting up in the morning and expressing their views on twitter while have their coffee and breakfast. Lowest number of tweets were tweeted during the hours 12am – 5am, which is understandable as most of the populace would be fast asleep during the night and early hours.

Finally, a word cloud was created which provided an insight into the most frequently used words in the tweets during this period effected by Covid-19.



to kickstart it's economy. A motivated populace would not only support its businesses but it will also would help in the curbing of this menace at a much swifter pace.

As from our project, we were able to see through the sentiment analysis that the people of the Republic of Ireland have been quite supportive and motivated to help in kickstarting the country's economy and also stopping the spread of the virus by working together. As a result, a flattening of the curve was seen which helped the country to eventually come out of lockdown.

I would feel that this type of analysis and results are quite scalable and highly reproducible and therefore this could be used to provide insights to important entities, government bodies and other stakeholders to devise people friendly decisions or policies. Of course, there have been a few limitations which are as follows:

- Data collected is from a specific time period and I believe more data can be collected from other sources to get more accurate results.
- Due to limitation of the python script, most tweets collected were from Dublin, hence the analysis might not give an overall picture of the mood of the country. Although, it is to be noted that close to 1.4 million people stay and work in Dublin.
- Another limitation or issue can be the ambiguous nature of some tweets which might not provide an accurate sentiment analysis.

Through this project, I was able to demonstrate a comprehensive understanding of the sentiment analysis through the CRISP-DM methodology. I was able to use and involve lot of different tools into this project, for example, MS Excel, Python, R programming, RapidMiner and PowerBI for the visualization.

## 5 APPENDIX

### 5.1 Appendix 1 – Python script for Data collection

```
#install and import module 'GetOldTweets3' and 'pandas'
import GetOldTweets3 as got
import pandas as pd

tweetCriteria = got.manager.TweetCriteria()

tweetCriteria.setQuerySearch("Covid19" or "covid" or "corona" or "stayathome" or "lockdown" or
"coronavirus") #keywords to search in the tweets
tweetCriteria.setSince("2020-03-01") #from date
tweetCriteria.setUntil("2020-07-15") #to date
tweetCriteria.setMaxTweets(10000) #max tweets to extract
tweetCriteria.setNear("Dublin") #location to get tweets from

tweets = got.manager.TweetManager.getTweets(tweetCriteria)
text_tweets = [[tweet.date, tweet.text, tweet.hashtags, tweet.retweets] for tweet in tweets]

#dataframe with columns 'Datetime', 'Text', 'Hashtag' and 'Retweets'
tweets_df = pd.DataFrame(text_tweets, columns=['Datetime', 'Text', 'Hashtag', 'Retweets'])

#writing the dataframe to a csv file
tweets_df.to_csv('E:\\DA Salil\\Final Project\\Tweet\\Dublin.csv')
quit()
```

## 5.2 Appendix 2 – R script to merge different .csv files

```
#listing the filenames to be merged.
filenames = list.files(path="E:\\DA Salil\\Final Project\\Tweet\\Tweets by county",pattern="*.csv")

#printing the filenames to be merged
print(filenames)

#full path name to the csv filenames
fullpath = file.path("E:\\DA Salil\\Final Project\\Tweet\\Tweets by county",filenames)

#from the path above, merge the listed files
senti = do.call("rbind",lapply(fullpath,FUN=function(files){ read.csv(files)}))

#print the merged csv dataset, using `head()` function to get first few rows of merged dataset
head(senti)

#writing the merged dataset to a csv file
write.csv(senti, "E:\\DA Salil\\Final Project\\Tweet\\Tweets by county\\onefile.csv",row.names=F)
```

### 5.3 Appendix 3 – Cleaning dataset and performing Sentimental analysis in R

```
library(sentimentr) #calculate text polarity sentiment at sentence level
library(tidyverse) #set of packages includes libraries for data visualisation, data manipulation
                    #and data tidying
#read csv file consisting of the raw data
covid = read.csv("~/Desktop/R csv/onefile2.csv", header = T, fileEncoding = "latin1")

#No. of tweets by location
#location = covid$Location
#loc.freq = table(location)
#barplot(loc.freq, col = rainbow(20))

#creating dataframe
covid.df = as.data.frame(covid)

#cleaning tweets with gsub() function

#text to lower case
covid.df$Text = tolower(covid.df$Text)

#replace http with blank
covid.df$Text = gsub('http\\S+\\s*', "", covid.df$Text)
covid.df$Text = gsub('https\\S+\\s*', "", covid.df$Text)
#View(covid.df$Text)

#replace punctuations with blank
covid.df$Text = gsub("[[:punct:]]", "", covid.df$Text)
```

```
#replace alphanumeric words with blank
covid.df$Text = gsub("[^0-9A-Za-z//' ]", "", covid.df$Text)
```

```
#replace digits with blank
covid.df$Text = gsub("\\d+", "", covid.df$Text)
```

```
#replace rt with blank
covid.df$Text = gsub("rt", "", covid.df$Text)
```

```
#remove @ with blank
covid.df$Text = gsub("@\\w+", "", covid.df$Text)
```

```
#replace common words with blanks
covid.df$Text = gsub("covid", "", covid.df$Text)
covid.df$Text = gsub("covid19", "", covid.df$Text)
covid.df$Text = gsub("corona", "", covid.df$Text)
covid.df$Text = gsub("coronavirus", "", covid.df$Text)
covid.df$Text = gsub("virus", "", covid.df$Text)
covid.df$Text = gsub("ireland", "", covid.df$Text)
```

```
#View(covid.df)
```

```
#convert blank cells to NA
covid.df$Text[covid.df$Text == " "] = NA
View(covid.df)
```

```
#omit NA
covid.df = na.omit(covid.df)
View(covid.df)
```

```
#sentiment analysis
covid_tw = get_sentences(covid.df$Text)
```

```
covid_tw = sentiment(covid_tw)
View(covid_tw)

#include a column 'sentiment' in dataframe
covid.df$sentiment = as.numeric(as.character(covid_tw$sentiment))
#View(covid.df)

hist(covid.df$sentiment, breaks = 20)

#label sentiments to Positive, Negative and Neutral
covid.df$mood = ifelse(covid.df$sentiment > 0, "Positive",
                      ifelse(covid.df$sentiment < 0, "Negative",
                              "Neutral"))
covid.df = na.omit(covid.df)
#View(covid.df)

#write labelled dataframe into csv file
write.csv(covid.df, file = "~/Desktop/R csv/labelled one file final.csv")
```

## 5.4 Appendix 4 – Word Cloud in R

```
library(tm)
library(textmineR)
library(wordcloud2)
library(tidytext)
library(textdata)
library(tidyverse)
library(textclean)

#Importing dataset
data = read.csv("~/Desktop/R csv/labelled one file final.csv")
#View(data)

#Converting dataset into a corpus
data_text <- iconv(enc2utf8(data$Text),sub="byte")
data_corpus = Corpus(VectorSource(data_text))
#inspect(data_corpus[1:5])

#Data pre-processing
#Convert all text to lower case
data_corpus <- tm_map(data_corpus, tolower)
#inspect(data_corpus[1:5])

#To plain text
```

```

data_corpus <- tm_map(data_corpus, PlainTextDocument)
#inspect(data_corpus[1:5])

#Punctuation removed
data_corpus <- tm_map(data_corpus, removePunctuation)
#inspect(data_corpus[1:5])
for (i in seq(data_corpus)) {
  data_corpus[[i]] <- gsub('[^a-zA-Z[:blank:]]', "", data_corpus[[i]])
}
#inspect(data_corpus[1:2])

#function for removing 'http' urls
removeURL<- function(x) gsub("http[[:alnum:]]*", "", x)
data_corpus<- tm_map(data_corpus,content_transformer(removeURL))

inspect(data_corpus[1:1])

#Remove stopwords words
data_corpus <- tm_map(data_corpus, removeWords, stopwords(kind='en'))
data_corpus <- tm_map(data_corpus, removeWords, c("covid", "coronavirus", "amp", "can", "will",
"now"
, "just", "a", "an", "the"))
inspect(data_corpus[1:1])

#remove non "American standard code for information interchange (curly quotes and ellipsis)"
# using function from package "textclean"

removeNonAscii<-function(x) textclean::replace_non_ascii(x)
data_corpus<-tm_map(data_corpus,content_transformer(removeNonAscii))

```

```
data_corpus<- tm_map(data_corpus,removeWords,c("amp","ufef",
      "ufeft","uufefuufefuufef","uufef","s", "for", "lot"
      , "much", "get"))
```

```
#strip whitespace
```

```
data_corpus<- tm_map(data_corpus,stripWhitespace)
```

```
#inspect(data_corpus[1:1])
```

```
#View(data_corpus)
```

```
#term document matrix
```

```
tdm = TermDocumentMatrix(data_corpus)
```

```
tdm
```

```
tdm = as.matrix(tdm)
```

```
#tdm[1:50,1:1]
```

```
w = rowSums(tdm)
```

```
#wordcloud2
```

```
w = data.frame(names(w),w)
```

```
colnames(w) = c("word", "freq")
```

```
wordcloud2(w, size = 0.5, shape = "pentagon",
```

```
  rotateRatio = 0.4, minSize = 4)
```

## 6 REFERENCES

- Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R. J. 'Sentiment analysis of twitter data'. 2011, 30-38.
- Chen, S. (2020) *Getting Started with Text Vectorization* Understand Natural Language Processing(NLP) — Text Vectorization in Python. Available at: <https://towardsdatascience.com/getting-started-with-text-vectorization-2f2efbec6685>.
- Core, R. S. (2020) *F<sub>ast</sub> Large Margin*. Available at: [https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/support\\_vector\\_machine\\_s/fast\\_large\\_margin.html#:~:text=The%20Fast%20Large%20Margin%20operator,Fan%2C%20K.W.&text=An%20SVM%20model%20is%20a,is%20as%20wide%20as%20possible](https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/support_vector_machine_s/fast_large_margin.html#:~:text=The%20Fast%20Large%20Margin%20operator,Fan%2C%20K.W.&text=An%20SVM%20model%20is%20a,is%20as%20wide%20as%20possible).
- Europe, S. V. *What is the CRISP-DM methodology?* Available at: <https://www.sv-europe.com/crisp-dm-methodology/>.
- Gov.ie (2020) *Government publishes roadmap to ease COVID-19 restrictions and reopen Ireland's society and economy*. Press Release. Available at: <https://www.gov.ie/en/press-release/e5e599-government-publishes-roadmap-to-ease-covid-19-restrictions-and-reopen/>.
- Gupta, P. (2017) *Cross-Validation in Machine Learning*. Available at: <https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f>.
- Hoare, T. (2020) *Text Mining*. Available at: [https://elearning.dbs.ie/pluginfile.php/1125275/mod\\_resource/content/0/Specialized%20Applications%20-%20Text%20Mining.pdf](https://elearning.dbs.ie/pluginfile.php/1125275/mod_resource/content/0/Specialized%20Applications%20-%20Text%20Mining.pdf).
- Kari Paul, D. R. (2020) *Tech giants' shares soar as companies benefit from Covid-19 pandemic*. Available at: <https://www.theguardian.com/business/2020/jul/30/amazon-apple-facebook-google-profits-earnings>.
- Kelly C De Bruin, E. M., Aykut Mert Yakut (2020) *The environmental and economic impacts of the COVID-19 crisis on the Irish economy: An application of the I3E model*. Available at: <https://www.esri.ie/publications/the-environmental-and-economic-impacts-of-the-covid-19-crisis-on-the-irish-economy-an>.
- Learn, M. (2020) *Sentiment Analysis*. Available at: <https://monkeylearn.com/sentiment-analysis/#:~:text=Sentiment%20analysis%20is%20the%20interpretation,in%20online%20conversations%20and%20feedback>.
- Lora Jones, D. P. D. B. (2020) *Coronavirus: A visual guide to the economic impact*. Available at: <https://www.bbc.com/news/business-51706225>.
- Nuggets, K. (2015) *Sentiment Analysis 101*. Available at: <https://www.kdnuggets.com/2015/12/sentiment-analysis-101.html#:~:text=Different%20types%20of%20sentiment%20analysis%20use%20different%20strategies%20and%20techniques,%2Faspect%2Dbased%20sentiment%20analysis>.
- Raschka, S. (2014) *Naive Bayes and Text Classification*. Available at: [https://sebastianraschka.com/Articles/2014\\_naive\\_bayes\\_1.html#n-grams](https://sebastianraschka.com/Articles/2014_naive_bayes_1.html#n-grams).
- Saini, S., Punhani, R., Bathla, R. and Shukla, V. (2019) *Sentiment Analysis on Twitter Data using R*.
- Sharma, A.** (2020) *Cross Validation in Machine Learning*. Available at: <https://www.geeksforgeeks.org/cross-validation-machine-learning/>.

Sidana, M. (2017) *Intro to types of classification algorithms in Machine Learning*. Available at: <https://medium.com/sifium/machine-learning-types-of-classification-9497bd4f2e14#:~:text=In%20machine%20learning%20and%20statistics,learning%20to%20classify%20new%20observations.&text=Linear%20Classifiers%3A%20Logistic%20Regression%2C%20Naive,Suport%20Vector%20Machines>.

TeleTradar (2020) *Ireland to go into two-week lockdown*. Available at: <https://www.teletrader.com/ireland-to-go-into-two-week-lockdown/news/details/51682688?internal=1&ts=1592954487093>.

Times, T. I. (2020) *Coronavirus timeline: The cases confirmed in Ireland so far*. Available at: <https://www.irishtimes.com/news/health/coronavirus-timeline-the-cases-confirmed-in-ireland-so-far-1.4195178>.

WHO Timeline - COVID-19 (2020). Available at: <https://www.who.int/news-room/detail/27-04-2020-who-timeline---covid-19>.

Wikipedia (2020) *Generalized linear model*. Available at: [https://en.wikipedia.org/wiki/Generalized\\_linear\\_model#:~:text=In%20statistics%2C%20the%20generalized%20linear,other%20than%20a%20normal%20distribution](https://en.wikipedia.org/wiki/Generalized_linear_model#:~:text=In%20statistics%2C%20the%20generalized%20linear,other%20than%20a%20normal%20distribution).

Worldometer COVID-19 CORONAVIRUS PANDEMIC. Available at: <https://www.worldometers.info/coronavirus/>.