National College of Ireland

BSc in Science(Honour) in Computing

Data Analysis

2021/2022

Lucas Villalba

X19120222

Lucasrv79@gmail.com

Irish Business Predictions by Twitter Data Mining

Technical Report

Contents

Executive Summary

The key point of this report is to find trends in social media communication(Tweets from Twitter) and company stock prices during the same period by analysing both social media and stock prices. Tweets drive stock, or the stock moves tweets.

There are indicators from the company performance side, such as revenue growth, revenue per client, profit margin, client retention rate, and customer satisfaction. Otherwise, this report will implement statistical government data from the Annual Business Survey of Economic Impact 2019 and private data as the current stock price of each company. Due to data exploration, The ABSEI 2019 is available from The Department of Business, Enterprise and innovation, part of the Irish government website offering access free to this novelty data resource in Excel readable format. With information divided by sectors and Irish owned companies with figures in euros from 2000 to 2019.

The private data, such as current stock price data, is also free from the Yahoo Finances website. Using the RStudio to extract and process this data comes in time-series form. The stock price by time with daily values such as open price, high price, low price, close price volume, and adjust—many potential uses for time-series.

On the Social media side, Twitter will be the target data source. The access and processing of data by RStudio, called Twitter Data Mining, will study patterns and trends concerning company performance from the social communication side. Natural Language Processing, which combines linguistics, computer science, and artificial intelligence, will be used as part of this process.

Contrasting and crossing the data above will tell us how sound companies perform on the stock market and how they drive their communication.

"...language is never innocent."

— Roland Barthes.

1.0 Introduction

1.1. Background

For the last thirty years, the relationship between the fields of Data Engineering and Business has grown at a fast and firm pace. This bond set the environment for e-commerce. Data engineering, a business resource, has undergone exponential growth (Beauchemin, 2017).

At the same time, Data Engineering and Business offer the whole Engineering and the entire business attributes from each to e-commerce reaches a global capability.

E-commerce is a field where companies want to hit the ground and run, offering exponential growth for those who drive well the data. The point of connection to the data analyst role is digging into data and finding accountable assets to enhance business.

For instance, it could be done by exploring historical data, like the company's stock prices with a timeline and applying regression to forecast or predict the future stock price o the company. Furthermore, we can extend our range of analysis with more data like company communication and see a correlation between stock price and company communication, how well they do, and how much they communicate it. On the other hand, how bad they do and how they communicate future implementation and corrections.

1.2. Aims

Find it there is a correlation between the emotions on tweets and the stock price, see if the stock price can be a modifier of company Tweets. On the other hand, Tweets can transform a company's stock price by emotions?

In either case, there is a pattern, a trend, or both in company tweets communication.

To context this first in economic background, concerning industries, then targeting the Bank of Ireland, Ryanair, and an external company like Agilent Technologies from the US, to see an outsider.

1.3. Technology

Twitter is an American microblogging, and social networking on which users post and interact with messages are known as tweets. Initially, with a limit of 140 characters, Textbase moved to 280 characters in 2017. they were founded in 2006 in San Francisco, California, US.

A Twitter developer account and creating a Twitter API application for the tweets extraction will need it. The documentation is used from the Twitter development (account, 2022).

RStudio is an integrated development environment IDE for R code and supports Python code, native (C/C++/Fortran) code open source and enterprise solutions. This includes a console syntax highlighting editor that supports direct code execution and a tool for plotting, history, debugging, and workspace management founded in 2009. Based on Boston, Massachusetts (RStudio, 2022).

1.4. Structure

- Data exploration and data wrangling.
- Visualisation, time-series plot and Sentiment analysis plot.
- Correlation and patterns.
- Statistical finding.
- Plotting results.
- Conclusion.

2.0 Data

During the module Software Project, we have many different presentations from different Lectures. From this series, one in particular, "*Finding good data*," on 1st November 2021, Lecture: Michael Bradford. It was the kick start for the data exploration to picture the Irish economy.

The private or public sector could do this; generally, the private sector offers this as a service, and it is not free, and sometimes it grasps some corporative thought.

On the other hand, the Irish government, as the public sector, offers free access to this data. The first port was the government website open data (Reform, 2022)the following data file, Annual Business Survey of Economic Impact 2019 was founded during the research. The source of this data is the Department of Public Expenditure and Reform, a government open data statistical dependable to frame the economy in terms of business in Ireland.

The Annual Business Survey of Economic Impact 2019 is a dataset from the Department of Enterprise, Trade and Employment; the dataset is in a CSV format file. It consists of five sheets: Summary, List of tables, All Agencies Client Companies, Irish Owned Companies, and Foreign Owned Companies.

Were Summary contents are three tables with fifty-one observations.

The sheet list of tables contents are the following:

- Table 1: Total Sales of Goods Produced and Services Provided (including sales to the dairy board)
- Table 2: Total Employment
- Table 3: Total Sales per Person Employed in Manufacturing and Int. Traded Services
- Table 4: Total Exports
- Table 5: Total Sales to the dairy board
- Table 6: Total Exports plus Sales to the Dairy Board
- Table 7: Total Exports as a percentage of Total Sales in Manufacturing and Int. Traded Services
- Table 8: Total Exports including sales to the Dairy Board as a percentage of Total Sales in Manufacturing and Int. Traded Services
- Table 9: Total Exports to the United Kingdom
- Table 10: Total Exports to Rest of Europe
- Table 11: Total Exports to Rest of Europe, EU Only (Only valid from 2013)
- Table 12: Total Exports to Rest of Europe, Non-EU (Only valid from 2013)
- Table 13: Total Exports to North America (figures prior to 2014 based on 2014 proportions)
- Table 14: Total Exports to Asia (figures prior to 2010 based on 2010 proportions)
- Table 15: Total Exports to Rest of World
- Table 16: Total Payroll Costs
- Table 17: Total Payroll Costs per Person Employed in Manufacturing and Int. Traded Services
- Table 18: Total Cost of Materials
- Table 19: Total Cost of Materials Sourced in Ireland
- Table 20: Irish-Sourced Materials as a Percentage of Total Materials
- Table 21: Total Cost of Services (including Royalties)

- Table 22: Total Cost of Services Sourced in Ireland (excluding Royalties Sourced in Ireland)
- Table 23: Irish-Sourced Services as a Percentage of Total Services
- Table 24: Total Cost of Services (excluding Royalties)
- Table 25: Irish-Sourced Services as a Percentage of Total Services (excluding Royalties)
- Table 26: Total Cost of Royalties
- Table 27: Total Cost of Royalties Sourced in Ireland
- Table 28: Total Value Added (Sales-(Total Materials Cost + Total Services Cost))
- Table 29: Total Value Added per person employed in Manufacturing and Int. Traded Services
- Table 30: Total Payroll Bill as a percentage of Value Added
- Table 31: Total Profit Estimate (Value Added - (Total Payroll + All Other Expense Items))
- Table 32: Total Profit as a Percentage of Total Sales
- Table 33: Total Profit per Person Employed in Manufacturing and Int. Traded Services
- Table 34: Proportion of Firms with any expenditure on Research and Development
- Table 35: Proportion of Firms with Total Expenditure on R&D >€127k Per Annum
- Table 36: Proportion of Firms with Total Expenditure on R&D >€250k Per Annum
- Table 37: Total Expenditure on Research and Development in Manufacturing and Int. Traded Services
- Table 38: Total Expenditure on R&D as a Percentage of Sales
- Table 39: Proportion of Firms with any expenditure on In-House Research and Development
- Table 40: Proportion of Firms with Total Expenditure on In_House R&D >€127k Per Annum
- Table 41: Proportion of Firms with Total Expenditure on In_House R&D >€250k Per Annum
- Table 42: Total Expenditure on In-House Research and Development in Manufacturing and Int. Traded Services
- Table 43: Total Expenditure on In-House R&D as a Percentage of Sales
- Table 44: Total Expenditure on In-House R&D Per Person Employed in Manu. And Int. Traded Services
- Table 45: Total Number of Persons Employed on In-House Research and Development
- Table 46: Total Number of Persons Employed on In-House R&D per thousand persons employed
- Table 47: Proportion of Firms with any expenditure on formal, structured training
- Table 48: Total Cost of all formal, structured training
- Table 49: Cost of all Structured, Formal Training as a percentage of payroll
- Table 50: Cost of all Structured Training per person employed
- Table 51: Direct Expenditure in the Irish Economy (Payroll, Irish Materials, Irish Services)

All the Agencies Client Companies, for each above, this sheet includes twenty variables covering the period from 2000 to 2020, with thirty-five observations as follows:

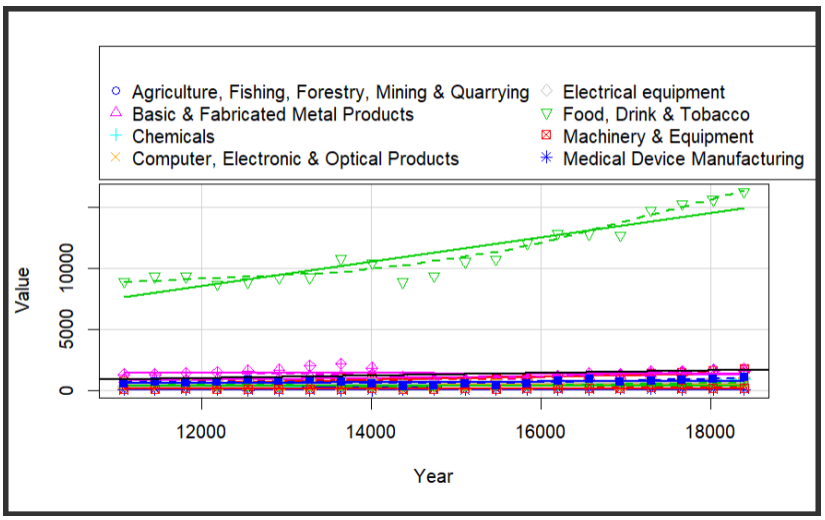
- Agriculture, Fishing, Forestry, Mining & Quarrying
- Food, Drink & Tobacco
- Textiles, Clothing, Footwear & Leather
- Wood & Wood Products
- Paper & Printing
- Chemicals
- Rubber & Plastics
- Non-Metalic Minerals
- Basic & Fabricated Metal Products
- Computer, Electronic & Optical Products
- Electrical equipment
- Machinery & Equipment
- Transport Equipment
- Medical Device Manufacturing
- Other Misc. Manufacturing
- Energy
- Recycling & Waste
- Construction (Excl. EI Amendments)
- Construction (EI Amendments)
- Publishing, Broadcasting & Telecommunications
- Computer Programming
- Computer Consultancy
- Computer Facilities Management
- Other IT & Computer Services
- Financial Services
- Business Services
- Education
- Other Services
- Dublin
- Rest of South & East
- BMW Area

Were Irish Owned Companies only shows Irish Owned Companies.

Then, the sheet of Foreign-Owned Companies shows only the companies with foreign owners. In order to visualise the groups over time and see the behaviour of each group, data wrangling was implemented to convert it from an excel file to RStudio as panel data.

The following workaround was made with the above data in RStudio. Please, for more information, check the following link http://dss.princeton.edu/training/Panel101.pdf . The panel data permits us to see the behaviour of each sector during the time from 2000 to 2019.

Where in Ireland, the sector “Food, Drink and Tabacco, has incredible growth. Please, see the following output on the next page, picture 2.1.



Picture 2.1

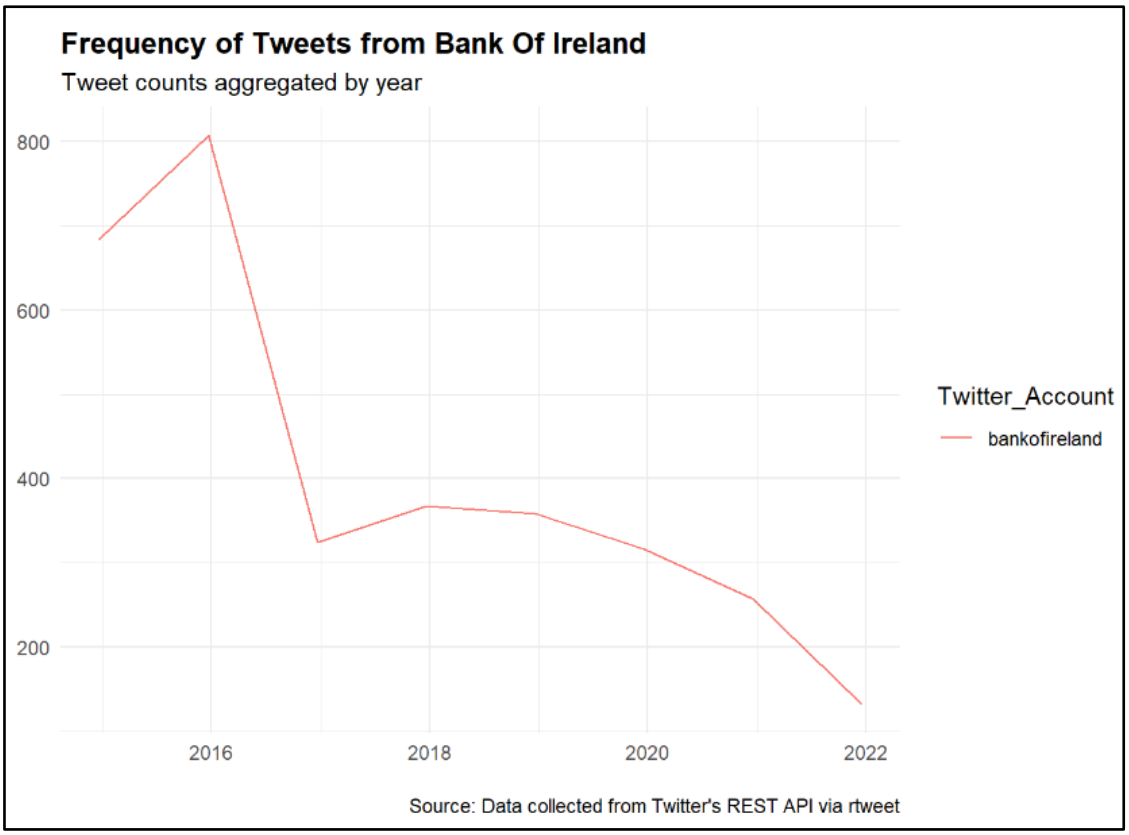
There is no missing data in the data elected.

A group of datasets from the Twitter API application, each dataset which thirty-thousand observations, each observation with ninety variables, the most used as follows:

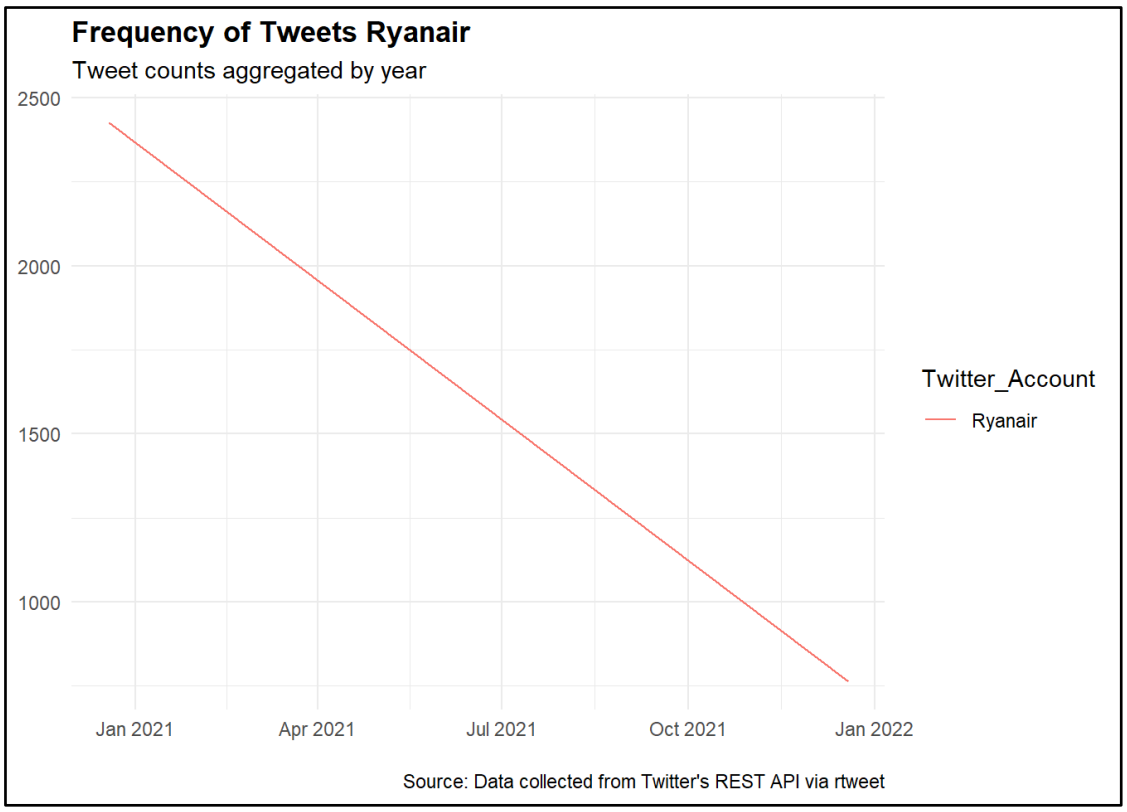
- <user_id> Twitter user id (character)
- <status_id> tweet status id (character)
- <created_at> date of the tweet (character)
- <screen_name> name targeted (character)
- <text> body of the tweet (character)
- <source> source of the tweet(character)
- <display_text_width> character width (numeric)
- <reply_to_status_id> alphanumeric value status_id (character)
- <replay_to_user_id> alphanumeric value user_id (character)
- <reply_to_screen_name> (character)
- <is_quote> true if it is a quote, false if it is not (boolean)
- <is_retweet> true if it is a retweet, false if it is not (boolean)

Natural Language Processing NLP is used to analyse a large amount of data, natural language, particularly words, as variables. NLP was used more in detail to analyse each Twitter account in terms of tweets, to have from those tweets outputs as the number of tweets for each company, to see more from text.

Different visualisation approaches were implemented inorder to ee tweets behaviour; the documentation was used for Twitter Data Mining (Rul, 2019). For each dataset, Bank of Ireland, Ryanair, and Agilent Technologies, initially, a frequency analysis was done for each company, where the number of tweets decreased significally, please see on the next page the picture 2.2 and the picture 2.3

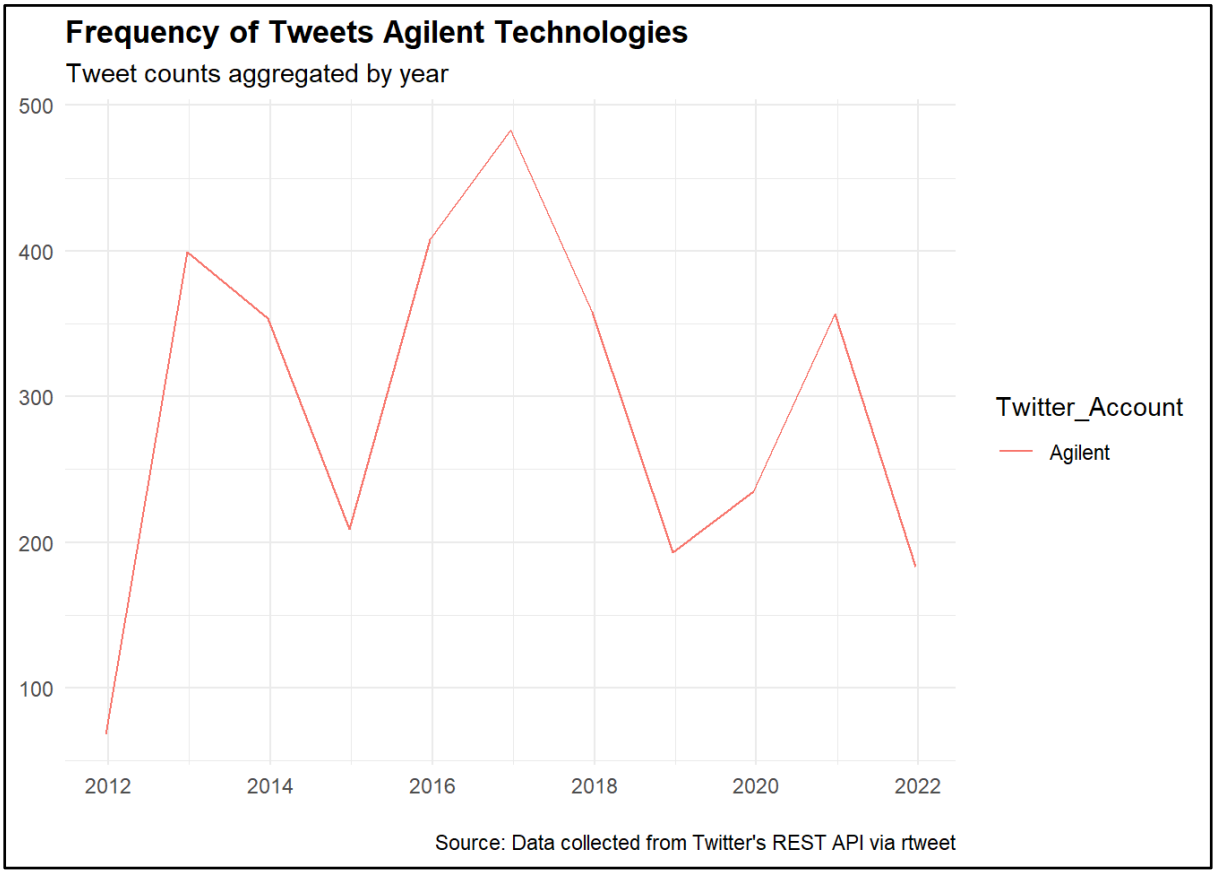


Picture 2.2



Picture 2.3

We take a US company into account to see if some difference is observed concerning the communication; Agilent Technologies' tweets frequency. Please, see picture 2.4.



Picture 2.4

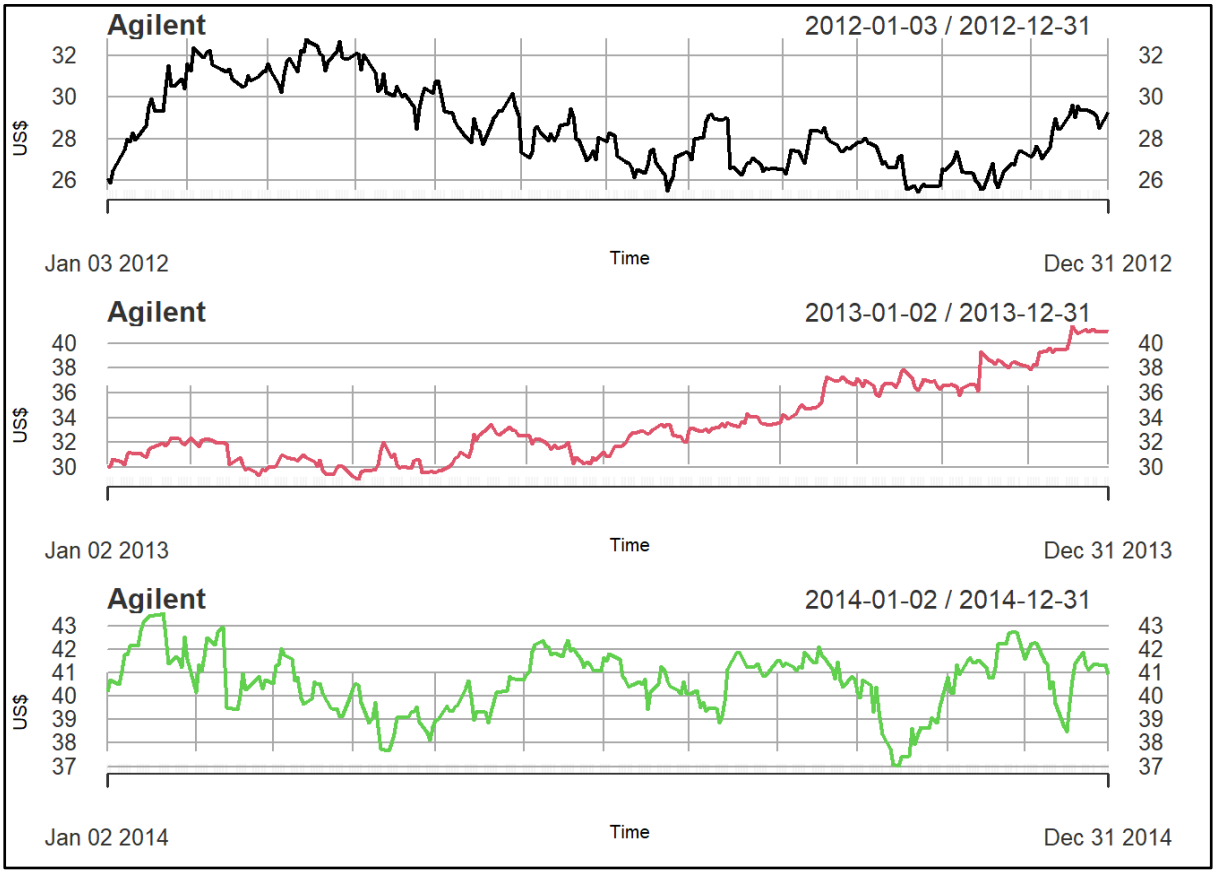
Finally, to context financial private data for each company, the stock price of each company data was required; Yahoo Finances offers a historical stock price dataset for any company in the stock market. The workaround was implemented using the <quadmod> library in RStudio (Data Camp, Joshua Ulrich, 2020).

The last group of datasets is a group of time series objects ready for the time-series process. Time-series allows visualising the stock price during a time. For each company of analysis, Bank of Ireland, Ryanair, and Agilent Technologies. each object with a time index with six variables in currencies such as Euro and American Dollars; this depends on the market, variables as follows:

- Open price
- High price
- Low Price
- Close Price

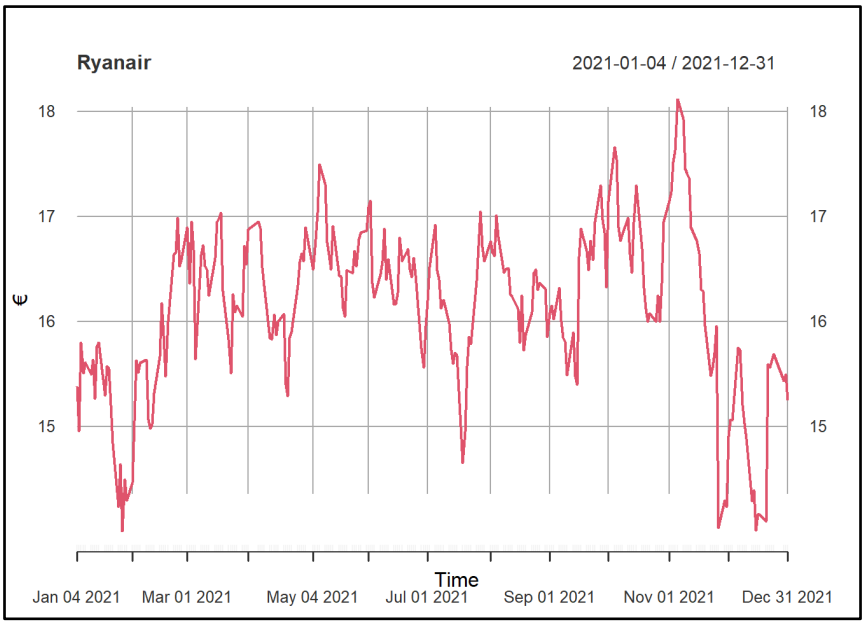
- Volume
- Adjusted

For the data exploration of the stock price data set, a single plot was implemented for each company to see the stock price fluctuation over time. The same company's stock price data was used for the following section in RStudio, which was used only for the close price in US$ for the years 2012, 2013, and 2014. Please, see the on the next page picture 2.5.



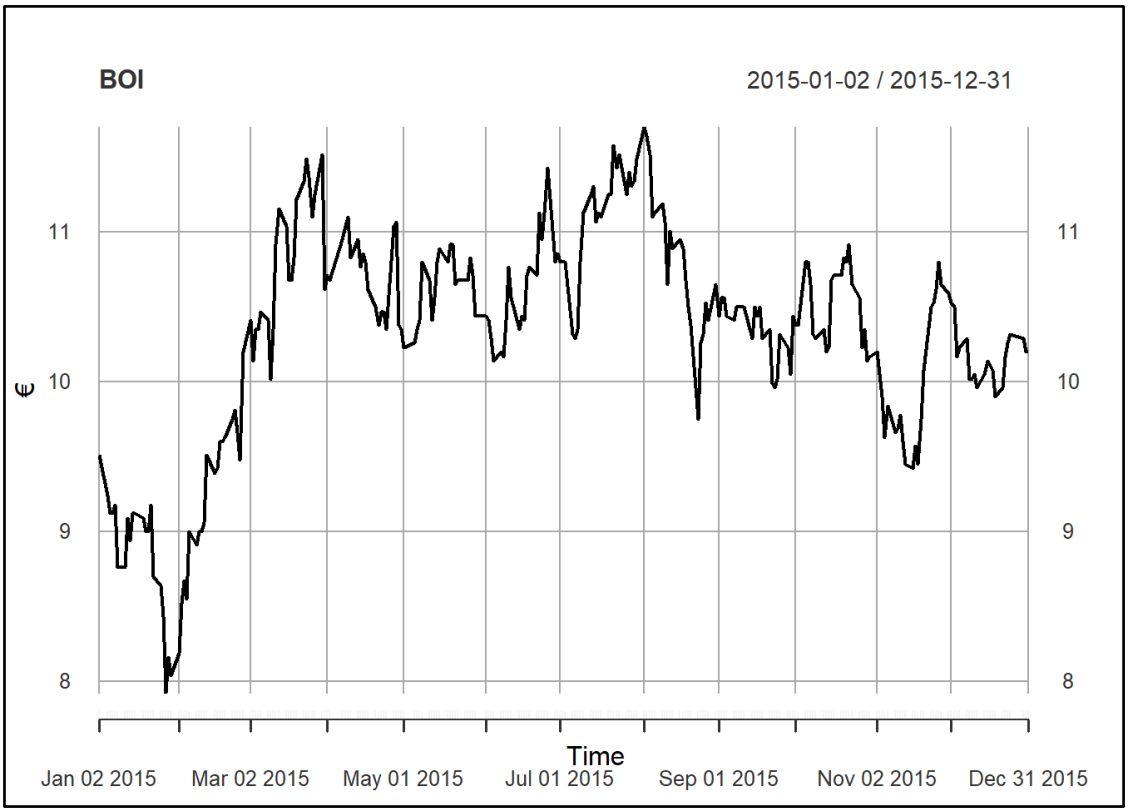
Picture 2.5

Also, the close stock price for the same period was conducted for Ryanair to see the fluctuation of the close price over time. We can identify a trend as the tweets decreased for 2021-2022. Please, see picture 2.6.



Picture 2.6

For the Bank of Ireland, to see the fluctuation of the close stock price, more in detail for 2015 because this sample year shows a peak in the Tweets numbers, we can identify a trend as the close price increase too for the same year. Please, see picture 2.7.

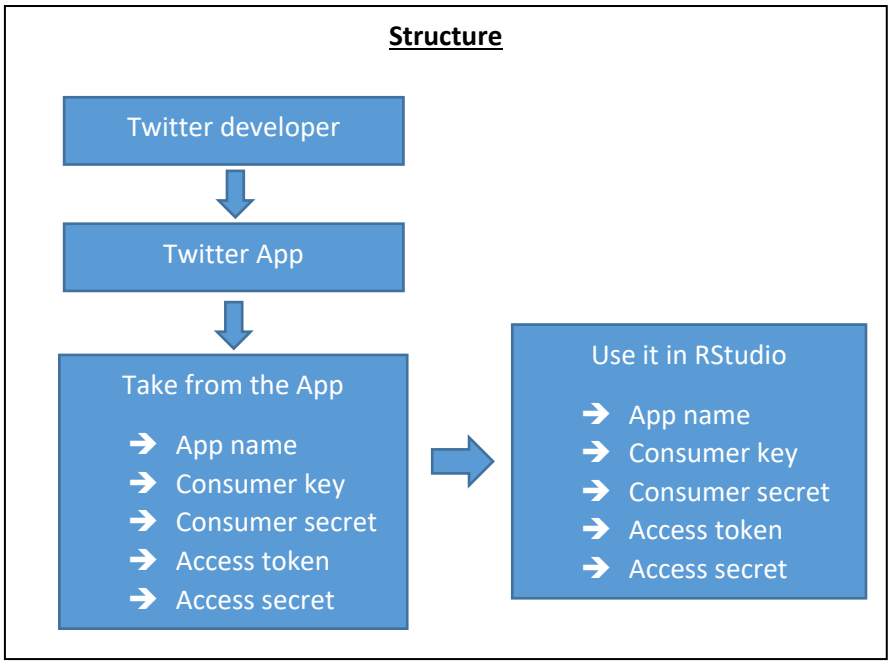


Picture 2.7

3.0 Methodology

As this report crosses industries, the methodology CRISP-DM stands for Cross-industry standard process for data mining method. It consists of six steps as follows; 1-Business understanding, 2-Data understanding, 3-Data preparation, 4-Modelling, 5-Evaluation, and 6-Deployment.

Initially, the workaround implemented for the Twitter data extraction was to create a Twitter developer account to have access via Twitter REST/API to extract the datasets. The following process is to create a Twitter App following the documentation of Twitter developer account support (account, 2022), with the following structure. Please see picture 3



Picture3

After creating the Twitter App, the same attributes will be used in the RStudio IDE with the (rtweet) package installed. Please see on the Suplementary Materia used section the code from RStudio for Twitter.

After this step is completed, we can proceed with the Twitter Rest API data extraction by using the <get_timelines> function. This function is part of the package<rtweet>, which allows extracting of data via stream API and REST. For the <get_timelines> we have parameters as:

- “user -> Vector of user names, user IDs, or a mixture of both.
- n -> number of tweets to return per timeline. Defaults to 100. Must be of length one or equal to the length of the user. This number should not exceed 3200 as Twitter limits return to the most recent 3,200 statuses posted or retweeted by each user.
- max_id -> Character, returns results with an ID less than (that is, older than) or equal to `max_id`.

- home -> Logical, indicating whether to return a user-timeline or home-timeline. By default, home is set to FALSE, which means get_timeline returns tweets posted by the given user. To return a user's home timeline feed, that is, the tweets posted by accounts followed by a user, set home to TRUE.
- parse -> Logical, indicating whether to return parsed (data.frames) or nested list object. By default, parse = TRUE saves users from the time [and frustrations] associated with disentangling the Twitter API return objects.
- check -> Logical indicating whether to remove check available rate limit. Ensures the request does not exceed the maximum remaining number of calls. Defaults to TRUE.
- token -> Every user should have their own Oauth (Twitter API) token. By default token = NULL this function looks for the path to a saved Twitter token via environment variables (which is what `create_token()` sets up by default during initial token creation). For instructions on how to create a Twitter token see the tokens vignette, i.e., `vignettes("auth", "rtweet")` or see ?tokens." (Camp, 2020)

Please, see picture 5 as an example:

Data extraction

```
# extract Bank of Ireland tweets
boi <- get_timelines("@bankofireland",
                     n = 10000,
                     language = "en",
                     since = "2014-01-01",
                     until = "2022-04-06",
                     token = twitter_token)
# write a CSV
write_as_csv(boi, "C:/Users/Administrador/Documents/R/Final year Project/Final Year Project/data/boi.csv", prepend_ids = TRU
E, fileEncoding = "UTF-8")
```

Picture 5

The above example shows the value n = "@companytwitteraccount" because, during the documentation review on a similar report, a different approach was discovered: tagging the n value with "@CEOofCompanyName".

Furthermore, the paper (Smith S, O'Hare A, 2022) pointed out the company's CEO tagging personal Twitter accounts due to people(CEO) facing business actions and reactions in most cases. This paper's conclusion shows there is no firm evidence of a correlation between tweets to stock price or vice versa. Additionally, there is minimal evidence, but this relies on an uncertainty zone due to censorship or filtering of tweets that prove tweets can drive stock price.

The second review on a similar report (Tahir M. Nisar, Man Yeung, 2017)states that the short period analysis is a crucial limitation due to the number of tweets. The case study used sixty thousand tweets per six days period of time. There is a reality that tweets are updated a hundred million per day.

To conclude our research with the following outputs for our analysis, first tagging companies rather than CEOs, second a prolonged study period, rather than analysed short period.

For the stock price dataset, Yahoo Finance was discovered, this is a free and time-series ready data object, and it can be accessed via RStudio IDE by installing the <quantmod> package. Please, see on the Suplementary Material used section the code from RStudio for Yahoo Finance.

Finally, the data from the Irish government, the Annual Business Survey of Economic Impact 2020, was manually downloaded. This data set is free and available at the government website, as part of open data project (Reform, 2022).

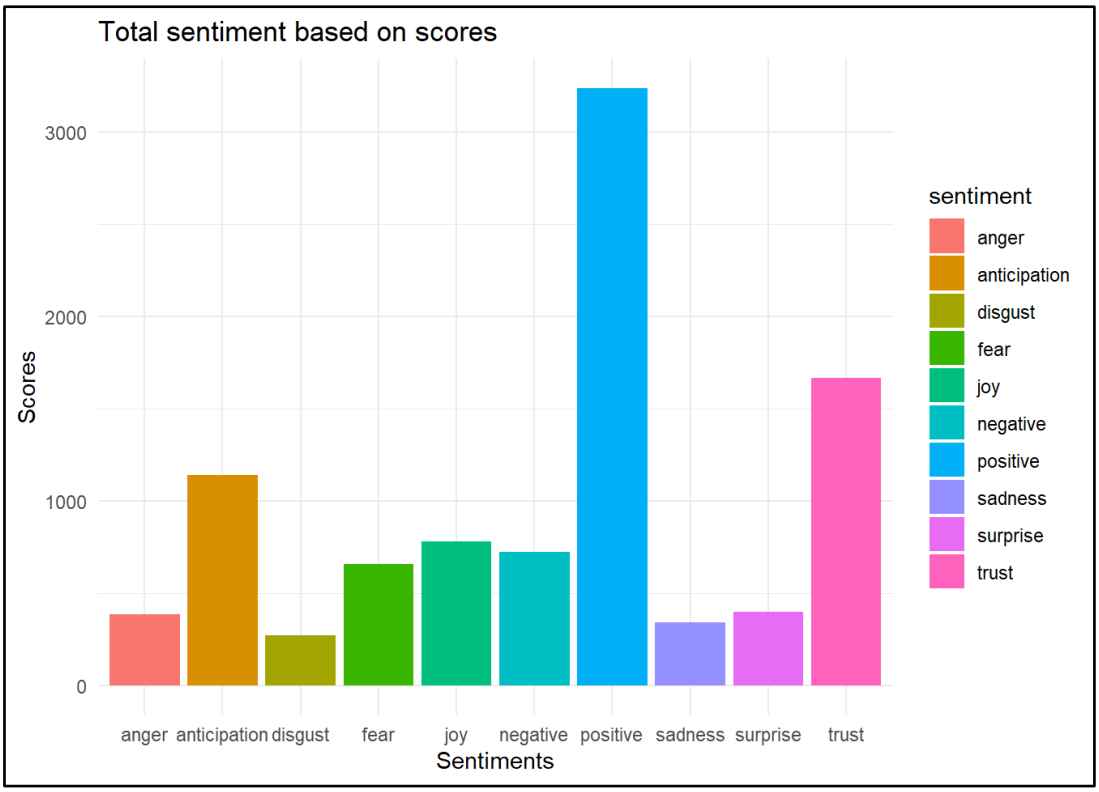
4.0 Analysis

Concerning the approach selected, was observed a significant difference in the score values for the Bank of Ireland for the 2015, and sentiment analysis was carried out with a score on Tweets' emotions for each of the three companies.

The <syuzhet> package offers the function <get_nrc_sentiment>, which analyses eight human emotions such as "*anger, fear, anticipation, trust, surprise, sadness, joy, and disgust and two sentiments negative and positive"* (Mohammad, 2022) *S*coring values in each emotion by four dictionaries.

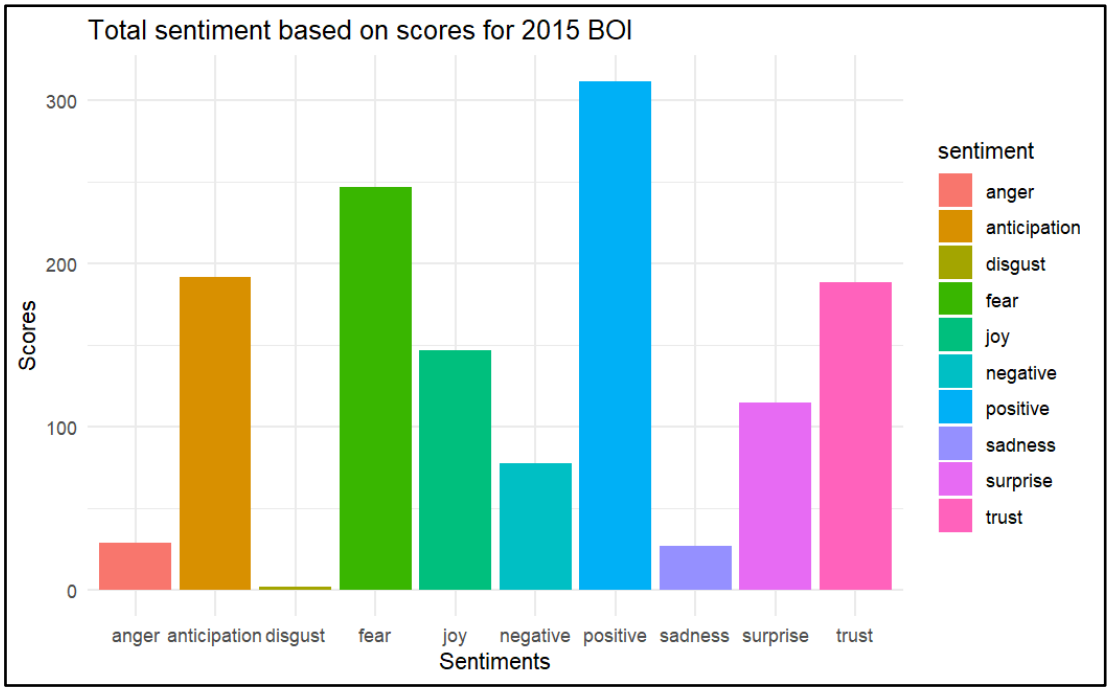
The author marks the importance of interpreting and understanding emotion into language for future Artificial Intelligence works.

The score for Agilent positive emotions passed the three-thousand figure, then trust emotion score with a value above one thousand five hundred, and finally the anticipation emotion score above one thousand. Please, see picture 6 on the next page.



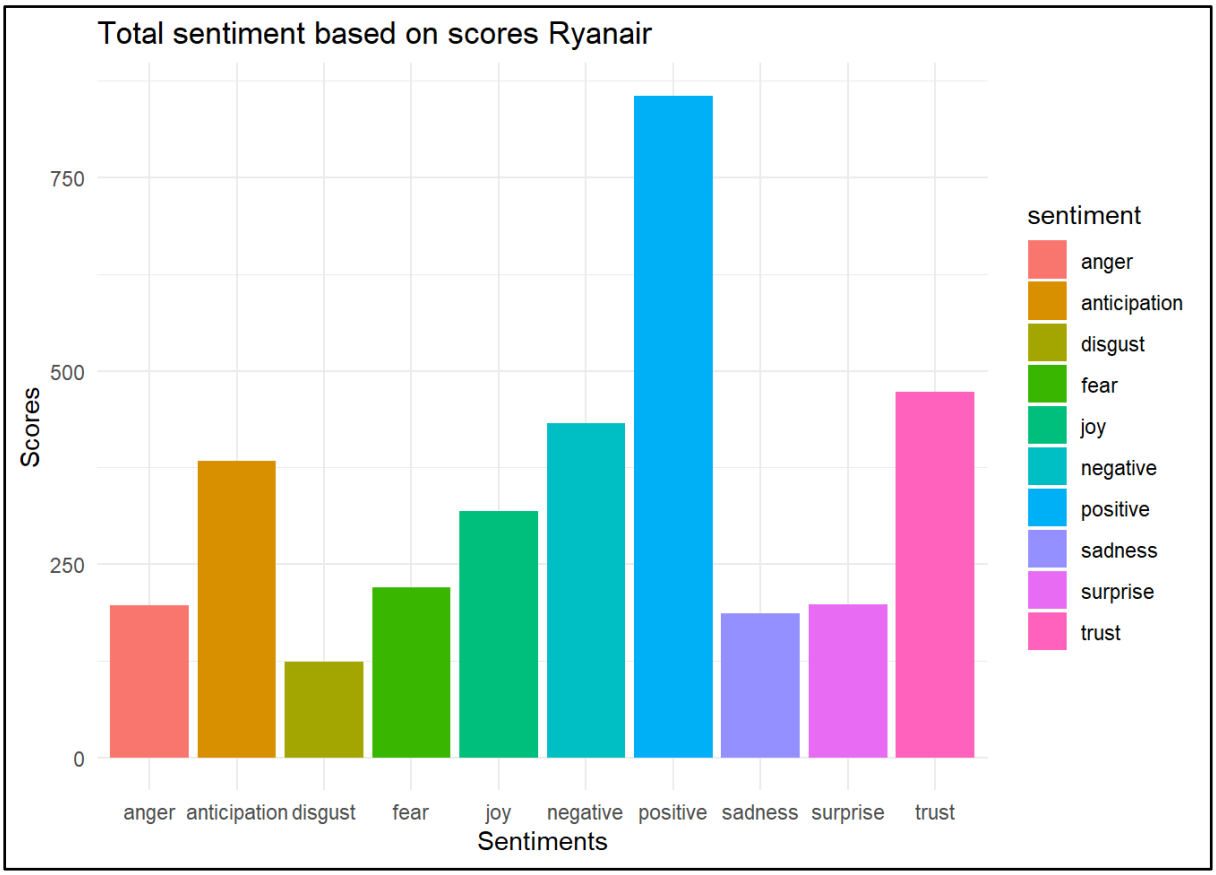
Picture 6.

For the Bank of Ireland, the time length was the same, but the figures were a bit lower than Agilent Technologies, with a positive score above one thousand fifty figure, the trust score above six hundred, and anticipation with a score above four hundred fifty. Please, see picture 7 on the next page.



Picture 7

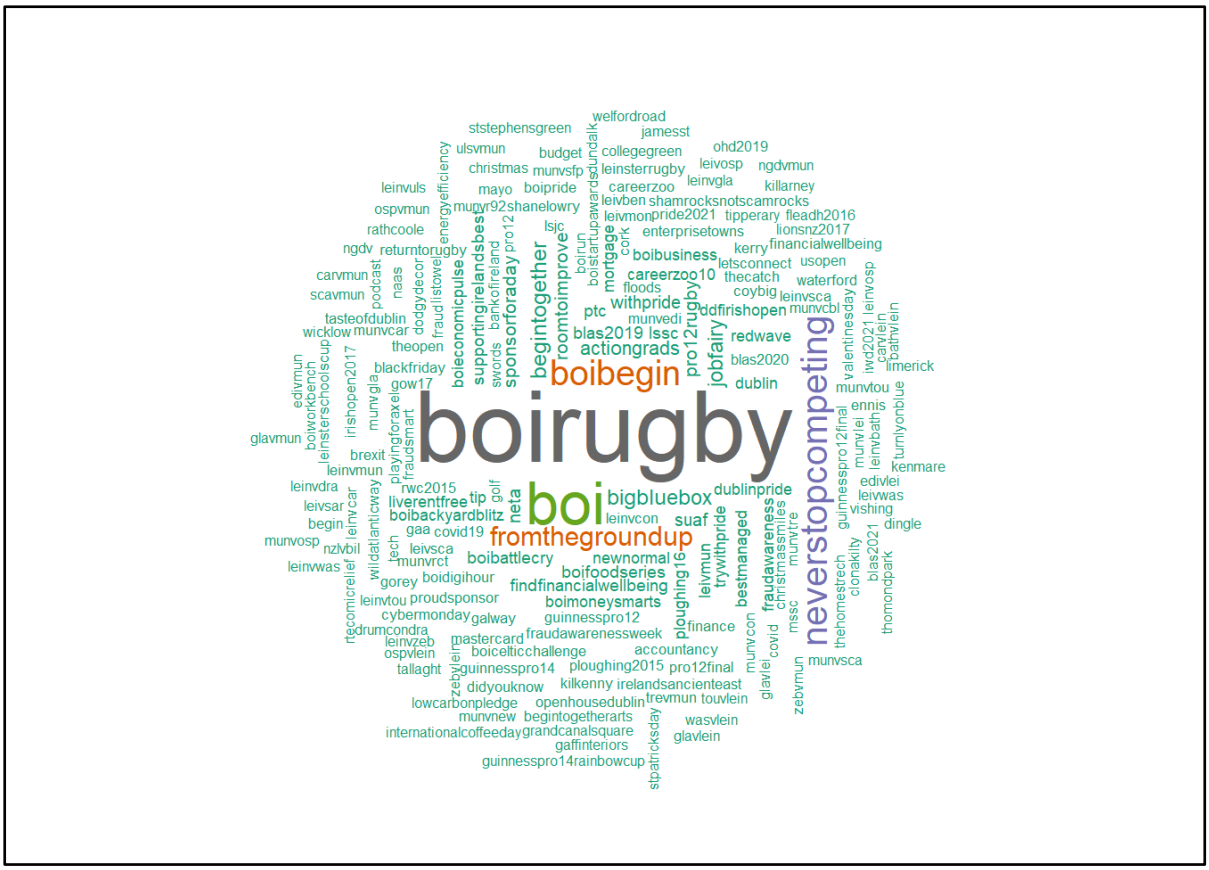
For Ryanair, a bit lower than Bank of Ireland, with a positive score under the figure of eight hundred seventy-five, the trust score under the figure of five hundred, and the trust with a value on the score of three hundred seventy-five. Please, see the picture 8 on the next page.



Picture 8

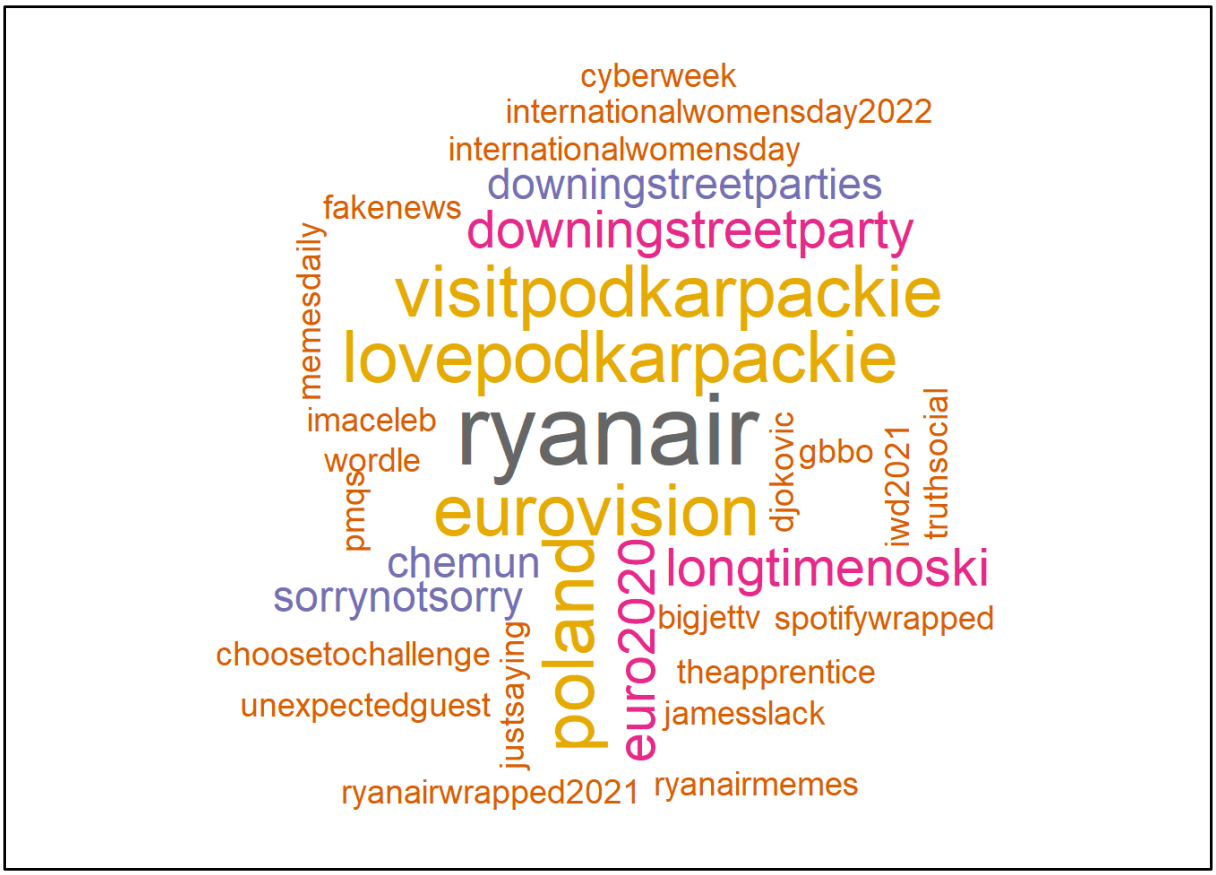
Further analysis was carried out about the what are the most common hashtags that each company posts in the Tweets, particularly in the organic Tweets, taking out retweets and replies from the analysis.

For the Bank of Ireland, the most common hashtags, <wordcloud> function to plot just the words before the hashtags for each vector of tweets for each company. Were in BOI, was founded “boirugby” was the most used. Please see picture 9 on the next page.



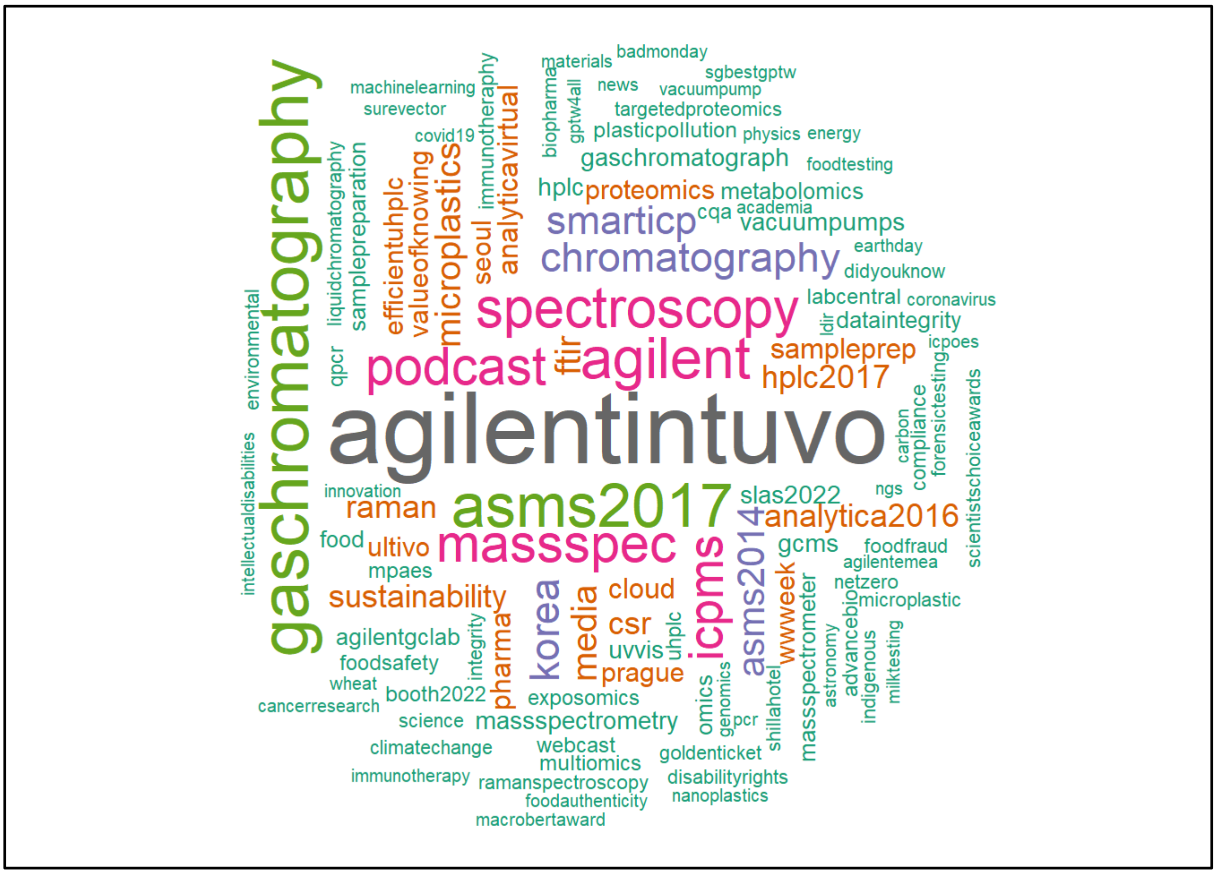
Picture 9

For Ryanair, the most common hashtags where the same name company was founded as the most used. Please see picture 10 on the next page.



Picture 10

For Agilent Technologies, the most common hashtag found agilentintuvo, is the start product from Agilent, which shows that the company communicates more about its development. Please see picture 11 on the next page.

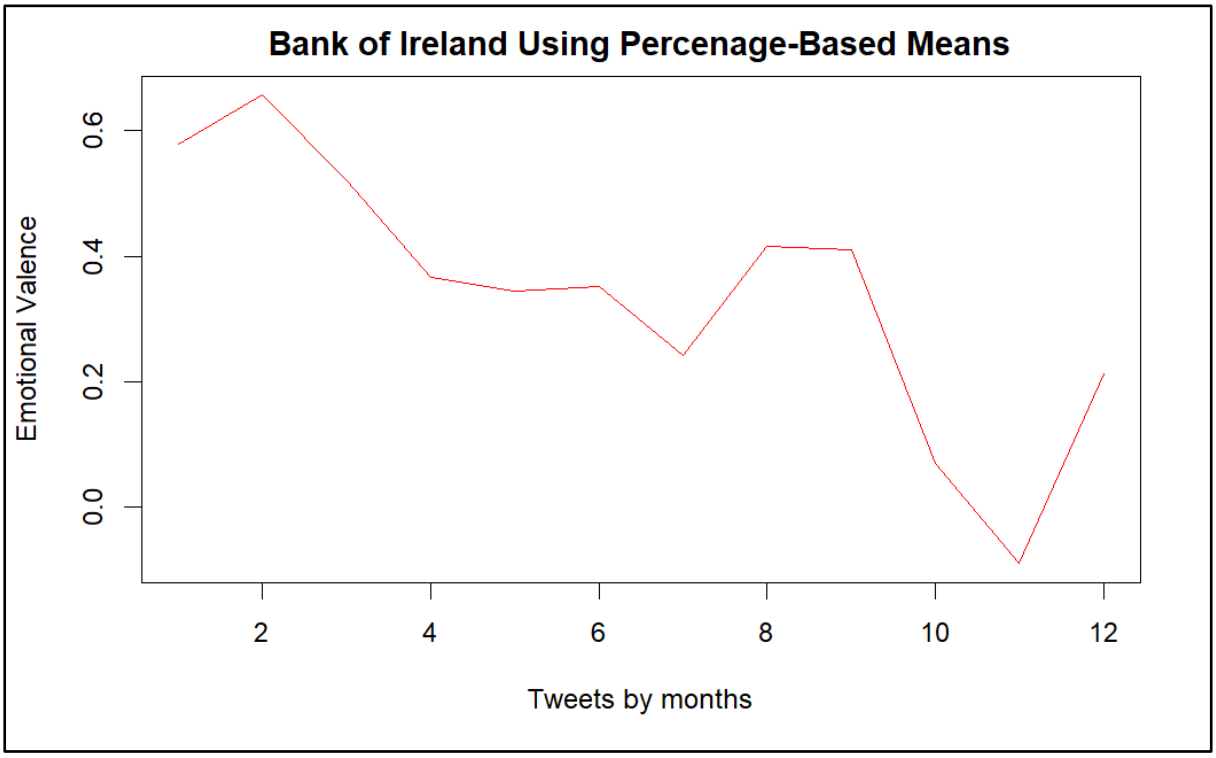

agilentintuvo
gaschromatography
spectroscopy
agilent
podcast
asms2017
massspec
chromatography
smarticp
icpms
korea
media
asms2014
microplastics
ftir
sustainability
raman
ultivo
analytica2016
hplc2017
sampleprep
proteomics
metabolomics
gaschromatograph
vacuumpumps
hplc
cloud
csr
prague
pharma
wwweek
massspectrometry
valueofknowing
efficientuhplc
seoul
analyticavirtual
immunotherapy
samplepreparation
liquidchromatography
machinelearning
surevector
covid19
materials
badmonday
sgbestgptw
news
vacuumpump
targetedproteomics
plasticpollution
physics
energy
foodtesting
biopharma
gptw4all
cqa
academia
earthday
didyouknow
labcentral
coronavirus
dataintegrity
ldir
icpoes
carbon
compliance
forensictesting
scientistschoiceawards
ngs
slas2022
gcms
foodfraud
agilentemea
netzero
advancebio
microplastic
astronomy
indigenous
milktesting
massspectrometer
shillahotel
pcr
genomics
omics
uhplc
uvvis
exposomics
integrity
agilentgclab
foodsafety
wheat
cancerresearch
booth2022
science
climatechange
immunotherapy
webcast
multiomics
ramanspectroscopy
foodauthenticity
macrobertaward
goldenticket
disabilityrights
nanoplastics
innovation
food
mpaes
qpcr
environmental
intellectualdisabilities

Picture 11

5.0 Results

From the statistical point of view, the correlation was suitable for conducting the results. The alternative approach was about seeing scores in the sentiment via tweets and contrasting this communication with the actual close stock price at the same time.

Firstly Bank of Ireland showed a score emotion related to 2015 Tweets associated with the momentum of the stock value as the score for emotions positive was high, but fear emotion was the second high value and trust in the third position, where the Emotional valence for the entire year divide by months shows a decrease. Please, see the picture 12.



Picture 12

Secondly, Ryanair shows a close relation to the decrement of tweets and stock values.

Finally, the outsider, Agilent Technologies US company, took the point of discrepancy during 2012; the close stock price rose from 26US$ to 30US$, and the tweets peaked at 400 for the same year. In 2013 the tweets decreased to 350, but the close stock price raised ten units from 30US$ to 40US$. Finally, for 2014, the tweets again reduced to 200, but the close stock price goes from 40US$ to 41US$.

6.0 Conclusions

In conclusion, the communication is relatively driven by the company's momentum; some patterns were observed in Agilent from a US company, seen the fluctuation of tweets and the stock price, but the evidence is not strong enough to build a trend to predict the stock price.

Also, the Irish companies show decent communication but in less number, less communicative than the US company. This was a surprise in the research as communication is vital for overall business performance and calls for future research.

7.0 Further Development or Research

Fort future research and using a similar approach, cloud computing will be implemented to collect more Tweets from different companies and contrast the UK market against the US and the Irish to see and find patterns that can enhance Irish business communication.

8.0 References

References

account, T. d., 2022. *Developer account support.* [Online]
Available at: https://developer.twitter.com/en/support/twitter-api/developer-account
[Accessed 3 November 2021].

Beauchemin, M., 2017. *The Rise of the Data Engineer.* [Online]
Available at: https://www.freecodecamp.org/news/the-rise-of-the-data-engineer-91be18f1e603/
[Accessed 15 05 2022].

Camp, D., 2020. *RStudio Documentation.* [Online]
Available at: https://www.rdocumentation.org/packages/rtweet/versions/0.7.0/topics/get_timeline
[Accessed 11 April 2022].

Data Camp, Joshua Ulrich, 2020. *getSymbols.yahoo: Download OHLC Data From Yahoo Finance.* [Online]
Available at: https://www.rdocumentation.org/packages/quantmod/versions/0.4.18
[Accessed 23rd November 2021].

Mohammad, S. M., 2022. *Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis.* [Online]
Available at: https://arxiv.org/abs/2109.08256
[Accessed 13th May 2022].

Reform, D. o. P. E. a., 2022. *IRELAND'S OPEN DATA PORTAL.* [Online]
Available at: https://data.gov.ie/
[Accessed 3rd November 2021].

RStudio, 2022. *RStudio.* [Online]
Available at: https://www.rstudio.com/products/rstudio/
[Accessed 21st September 2021].

Rul, C. V. d., 2019. *Towards Data Science.* [Online]
Available at: https://towardsdatascience.com/a-guide-to-mining-and-analysing-tweets-with-r-2f56818fdd16
[Accessed 20 November 2021].

Smith S, O'Hare A, 2022. Comparing traditional news and social media with stock price movements; which comes first, the news or the price change?. *Journal of Big Data,* 9(28th April 2022), p. 20.

Tahir M. Nisar, Man Yeung, 2017. Twitter as a tool for forecasting stock market movements:. *The Journal of Finance and Data Science,* 4(9th February 2018), p. 19.

9.0 Appendices

9.1. Project Proposal

9.2. Ethics Approval Application (only if required)

9.3. Reflective Journals

Supervision & Reflection Template

Student Name	Lucas Villalba
Student Number	X19120222
Course	Data Analyst

Month: October

What?

Plan to use a different dataset after the meeting today with Micahel B. The data from social networks will be due to project deliverables.

So What?

Success:

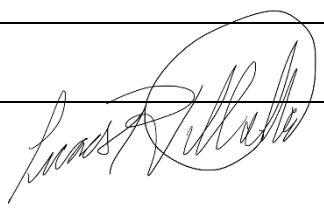
- Informatica account settled up. I used the college email account to set up a student account with Informatica, and access to a wide variety of software, tools were provided.

Challenge:

- The data from social networks will be more than an issue to work within this project.

Now What?

I will see different options in relation to data to use in the project.

Student Signature	

Supervision & Reflection Template

Student Name	Lucas Villalba
Student Number	X19120222
Course	Data Analyst

Month: November

What?

Plan to use a dataset from Gov.ie I am analysing the Datasets.

So What?

Success:

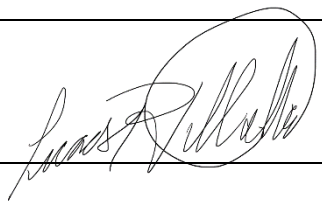
- Azure for VM signed up with 100 credits

Challenge:

- Time constraints related to bottleneck situation, other modules workload deadlines.

Now What?

Managed work, I have a second background that pays the rent and bills to reduce the quantity of time I spent at work by plus two more hours free per week.

Student Signature 	Date: 01/12/2021

Supervision & Reflection Template

Student Name	Lucas Villalba
Student Number	X19120222
Course	Data Analyst

Month: December

What?

The data that will be used as formal data is the Annual Business Survey of Economic Impact 2016 - DETE (enterprise.gov.ie)
The data to cross with the formal data above is the next task.

So What?

Success:

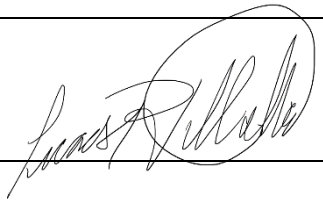
- Regression test done and set a Twitter account ready to track business twits related to Irish Business

Challenge:

- Do a sentiment analysis with the informal data collected from Twitter

Now What?

Researching in different approaches for sentiment analysis with a Twitter account.

Student Signature 	Date: 03/01/2022

Supervision & Reflection Template

Student Name	Lucas Villalba
Student Number	X19120222
Course	Data Analyst

Month: January 2022

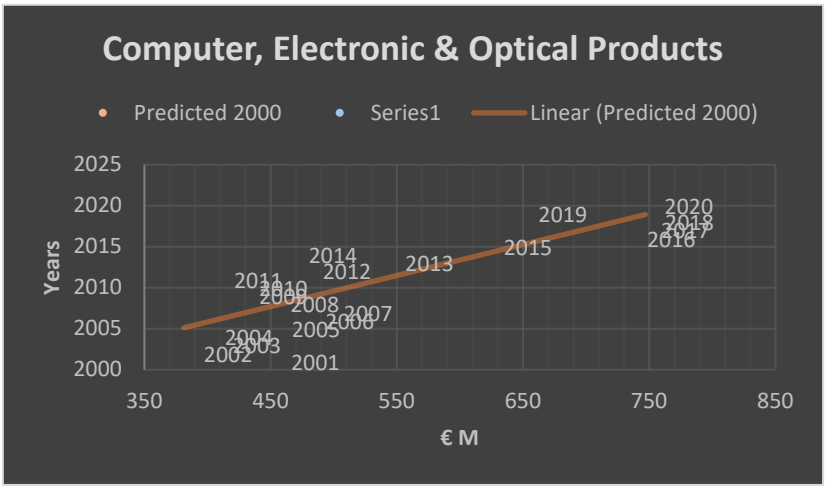
What? Formal Data crossed with informal data successfully, and there is a solid output.
Sentiment analysis was carried out in Rstudion, with the twits from the stock market 2019(informal data)

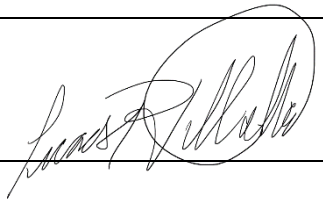
Text cloud twits from stock Market in 2019-Technology sector

Where Agilent Technologies is an American based company, with headquarters in Ireland,Cork.

~~

A regression analysis with moving average was carried out in Excel with the [Annual Business Survey of Economic Impact 2016 - DETE (enterprise.gov.ie)](#) (Formal data)



So What? Success: tere is a correlation in the technological sector that affect indirectly to the Tech-Irish owned companies, even if the sc Challenge: try to find domestic twits in relation to stock market or relative stock market twits, Euronext Dubin is not updated since a while, so alternative options as; Google Finances and different twitters from domestic stock market to research.	
Now What? Actively looking for twitters from Irish stock market.	
Student Signature 	28/01/2022

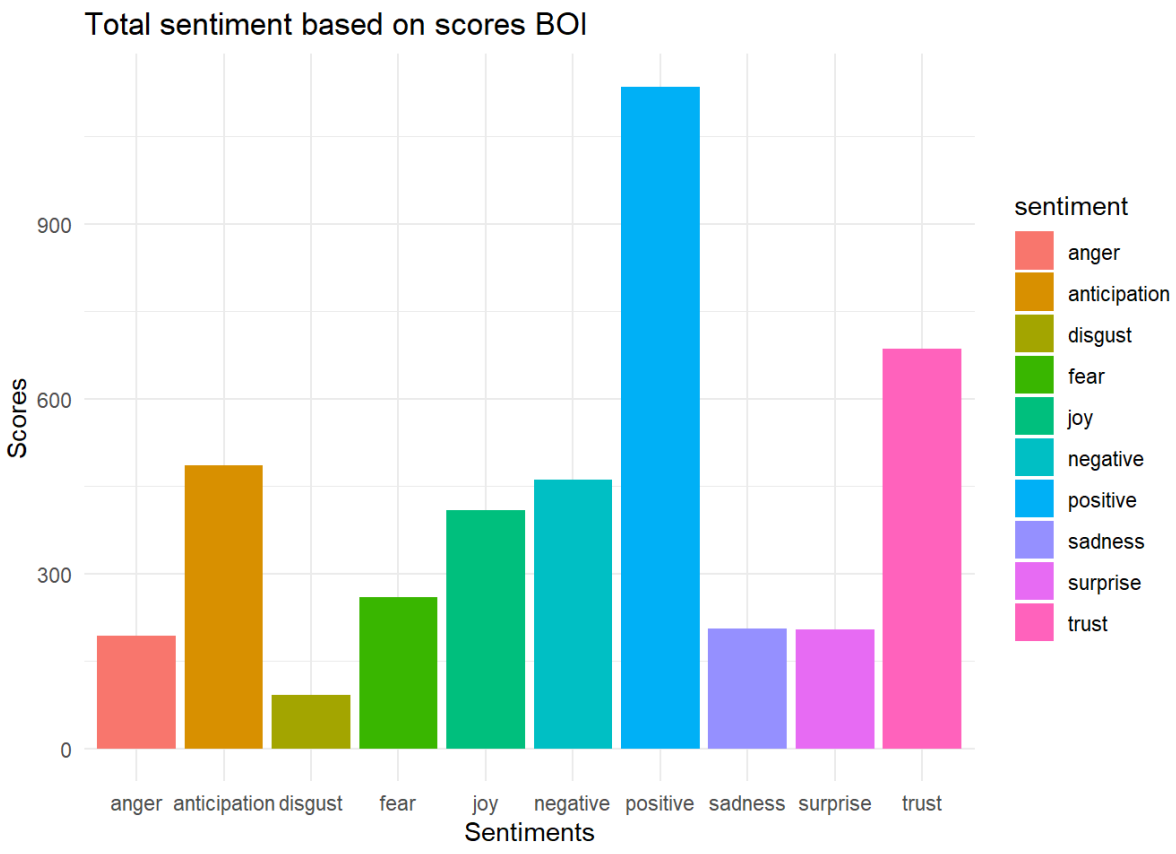
Supervision & Reflection Template

Student Name	Lucas Villalba
Student Number	X19120222
Course	Data Analyst

Month: February

What?

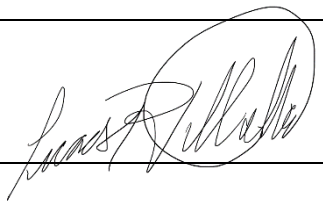
Found new documentation refers to the syuzhet package for RStudio in terms of Text Analysis using NPL Standford, which allows scoring eight human emotions plus negative and positive score words; this is the output for the Bank of Ireland about organic tweets.



So What?

Success:

- Ready to explore this package.

Challenge: • Looking for a data set clean, public access, and up-to-date for each company concerning some measure on the company assets.	
Now What? Lookin on Google Finance	
Student Signature 	28/02/2022

Supervision & Reflection Template

Student Name	Lucas Villalba
Student Number	X19120222
Course	Data Analyst

Month: March

What?

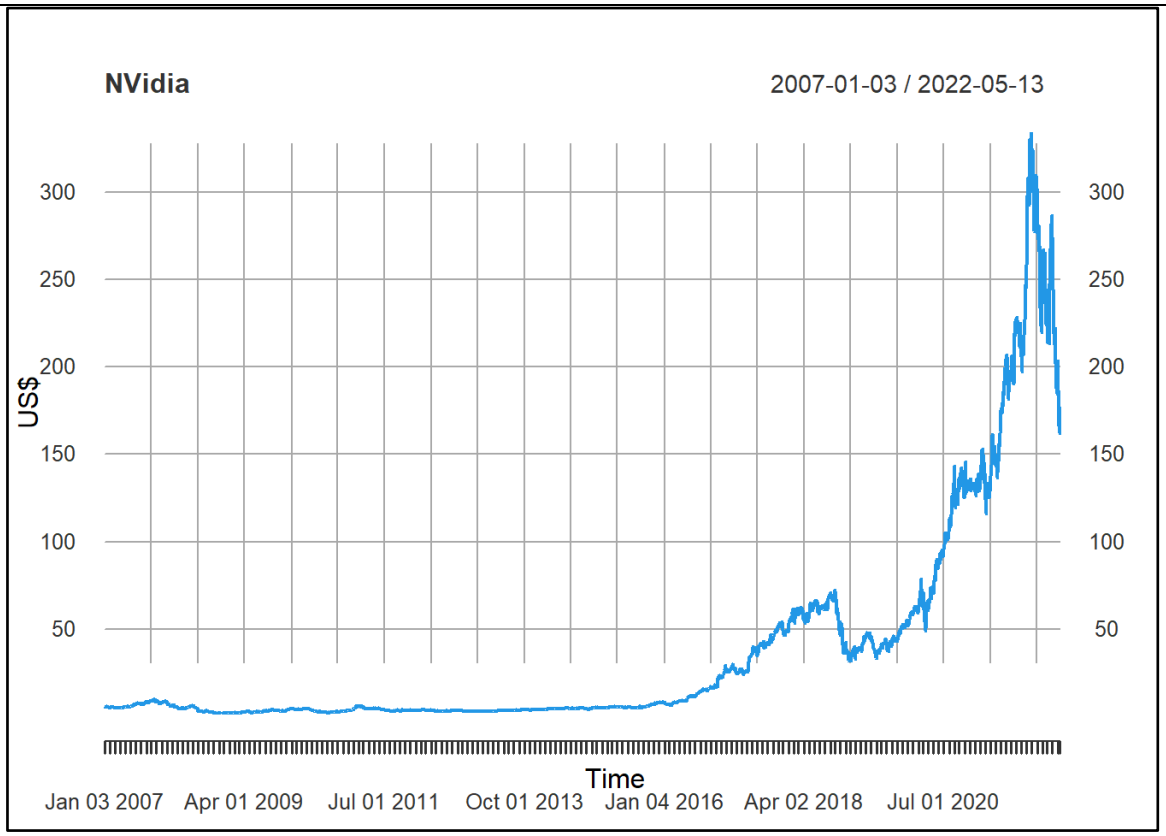
I found documentation related to extracting stock data from different companies in time-series format, which means it is historical data and can be processed directly. The extraction is by< quantmod> library, this by RStudio and call to the data by a symbol, which is the company's actual name into the stock market.

So What?

Success:

- Whit Quantmod, we can use the following sentence in RStuio

```
library(quantmod)

library(tseries)

library(timeSeries)

library(forecast)

library(xts)

library(Quandl)
symbol3 = "NVDA" # NVidia Holdings - Currency in US$

getSymbols(Symbols = symbol3, src = "yahoo")

#Plotting NVidia

par(mfcol = c(1, 1))

plot(NVDA[,4], major.ticks='months', minor.ticks =FALSE, main= "NVidia", col=4)



title(main = NULL,

    xlab = "Time",

    ylab = "US$")
```



Challenge:

- The next is to find a correlatio between the tweets and the stock price data

Now What?

Back to Statistical book notes

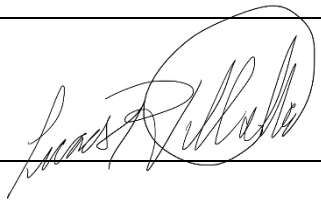
Student Signature	28/03/2022

Supervision & Reflection Template

Student Name	Lucas Villalba
Student Number	X19120222
Course	Data Analyst

Month: March

What? The correlation is not suitable for use between the tweets and stock price. The stock price value is daily, and the tweets are infrequent; this affects the hypotheses, and the visualisation is the alternative approach.
So What? Success: breaking down the historical stock price into years to see if there is any trend o pattern between the tweets and the stock price. Challenge: • Try to visualise more of the tweets and what we can get from them.
Now What? I am exploring text analysis Data Mining.

Student Signature 	28/04/2022

9.4. Suplementary materials used

RStudio Code for Twitter

```
# Whatever name you assigned to your created app
appname <- "yourAppName"
## api key (your key)
key <- "[redacted]"
## api secret (your secret)
secret <- "[redacted]"
# access_token
access_token <- "[redacted]"
#access token secret
access_secret <- [redacted]"

# Create token named "twitter_token"
twitter_token <- create_token(
  app = appname,
  consumer_key = key,
  consumer_secret = secret,
  access_token = access_token,
  access_secret = access_secret)
```

RStudio for Yahoo Finance.

```
library(quantmod)
symbol1 = "BIRG.IR" # Bank Of Ireland Group plc (BIRG.IR) Irish - Irish Delayed price. Currency in EUR
# Use getSymbols to import the data
getSymbols(Symbols = symbol1, src = "yahoo")

# Look at the first few rows of data of Bank Of Ireland Group plc (BIRG.IR) Irish and Class.
head(BIRG.IR)
class(BIRG.IR)
```

Were get_symbols “BIRG.IR” can be found in the Yahoo finance website, the search for news, symbols or companies section.