

National College of Ireland  
BSc in Computing  
2014/2015

Emlyn Farrell  
X14438218  
X14438218@Student.ncirl.ie

# A Statistical Analysis of Twitter Sentiment vs Share Price

Technical Report



# Table of Contents

<b>Executive Summary</b>	4
The solution to this is to compile three data sets	5
Sentiment Analysis	5
<b>Introduction</b>	6
Background	6
Aims	7
<b>Lit review</b>	7
Stock prediction using sentiment analysis	7
Does social media affect capital markets	8
Why do stock prices change based on news reports?	8
Understanding Sentiment Analysis: What It Is & Why It's Used	9
Data Mining	9
<b>Meeting With Journalist Frances Shanahan</b>	9
<b>Meeting with a Stockbroker Cathal O'Donnell (Cantor Fitzgerald)</b>	10
Main Points raised by Cathal	10
<b>Technologies I planned to use</b>	10
<b>Technologies used</b>	12
<b>System</b>	13
Requirements	13
Use case's	14
Requirement 1 <Creation & Populating of Datasets>	15
Requirement 2 <processing data>	17
Requirement 3 <Interpreting and exploring data>	19
Requirement 4 <Adding Datasets to Tableau & Rstudio >	21
Requirement 5 <Sentiment analysis>	23
Non-Functional Requirements	24
Performance/Response requirement	time 24 29

Design and Architecture	31
System Architecture	31
Implementation	31
Finding Datasets	31
Limiting what results are needed from each dataset	32
preprocessing of data	32
date formatting	32
Standardising date format	32
Filtering data	32
Sentiment analysis	33
Statistical tests	35
Breakdown of tweets by device	35
Breakdown of tweet sentiment	36
Test for normality	37
Descriptive Statistics	38
Correlation matrix	38
Graphical User Interface (GUI) Layout	41
Application Programming Interfaces (API)	41
Testing	41
User Testing	41
User testing table	42
Evaluation	46
<b>Results</b>	46
<b>Conclusions</b>	<b>Error! Bookmark not defined.</b>
<b>Further development or research</b>	47
Lit review	48
Survey	48
Sentiment analysis	48
Overall	48
<b>References</b>	49
<b>Appendix</b>	50

Project Proposal	50
Monthly reports	50
User testing sheets from users	50
<b>Objectives</b>	50
Background research	52
Technical Approach	53
Special resources required	54
Technical Details	54
Evaluation	55
Technical Details	55
Project Plan	56
Semester 1	56
Semester 2 Jan to March	56
Semester 2 April to June	56
Monthly Journals	57
Monthly reports	57
OCT	57
NOV	58
JAN	59
<b>My Achievements</b>	59
<b>My Reflection</b>	59
<b>Intended Changes</b>	59
MAR	60
<b>My Achievements</b>	60
<b>My Reflection</b>	60
<b>Intended Changes</b>	60
<b>Supervisor Meetings</b>	60
Other Material Used	60
Testing scripts	<b>Error! Bookmark not defined.</b>
User 1	<b>Error! Bookmark not defined.</b>

User 2

**Error! Bookmark not defined.**

User 3

**Error! Bookmark not defined.**

## Executive Summary

A Statistical Analysis of Donald Trump's tweets to see whether his influence on twitter could be an extrinsic value that influences the share prices of stock-price indexes. Some Technologies being used are R, Mysql & Tableau to determine whether he has an extrinsic value using his tweets, stock prices & headlines

### ***The solution to this is to compile three data sets***

- Share price
  - This is got from Finance.yahoo.com/
  - Gathered by looking for stock indexes and getting the historical chart information for each stock index as laid out by Cathal (stockbroker)
  - Some indexes which cathal told me to look at where as follows
    - DJIA: Dow Jones Industrial average
    - FTSE 100: Financial Times Stock Exchange 100 Index
    - S&P 500: Standard & Poor's 500
    - NASDAQ: National Association of Securities Dealers Automated Quotations
- Tweets
  - The plan was to originally pull information through R studio using twitterR, which is a package which interfaces with the twitter API , but as there was limitations with how many results could be pulled back as each time the command for returning n=3200 tweets it would return anything from 650 to 900. to address this I sourced this dataset from <http://www.trumptwitterarchive.com/archive> and set my data parameters to 01/01/2017 to 31/12/2017 giving me a proper dataset to use
  -
- News Headlines
  - This was gotten from Kaggle which listed all of its headlines from the last 15 years <https://www.kaggle.com/therohk/million-headlines>

## ***Sentiment Analysis***

- The original plan was to write my own Python script that will be written with my own word banks to determine different types of sentiment and result in a score anywhere either side of 0 i.e. -5 to 0 to +5
- The way I actually carried it out was using a package within r called SentimentR

### Strengths of sentiment analysis

- Allows companies to measure public opinion

### Weaknesses of sentiment analysis

- It is unable to detect sarcasm, this is especially important in some parts of world especially in Ireland. An example being if a train is delayed a person might tweet "Great my train is late thanks Irish Rail for making my day better!", Sentiment analysis would rank this as a positive when in reality it is a false negative

## **Representing results**

- For this I had decided to build a dashboard in Tableau to visualise the results of the project, Tableau allows for the data to be manipulated in a user friendly environment so that user can find insight
- I will also be performing some statistical tests in Rstudio as well examples being

## **End Result**

- Is the question as set out in the introduction answered as a result of the Study that has been carried out
- To prove or disprove the hypothesis of the project title is there a difference or isn't there, can we find a correlation between the three datasets

## **● Introduction**

### ***Background***

- I've been thinking a lot about doing an Analytical type project for my final year, this was leveraged by work placement coordinator during my time with TQS Integration as it is something that would stand out in an interview, as opposed to the normal way of developing an App etc..
- The precursor to this is my avid interest in the stock market especially how it works and is able to influence the world economy.

- I've always loved how Twitter is always in the news and the way it is used to gauge of opinion, even if Twitter's usage is still quiet niche in terms of total active users and the demographics of people who use Twitter.
- I came up with the project idea around the time of when North Korea were threatening the US territory of Guam as I noticed during that week Donald Trump's twitter was being quite forward and convincing the world that we were on the brink of nuclear war, I noticed that during this period that the share prices of certain aerospace companies share prices increased dramatically during this period as well as other types of industries that supply the aerospace industry.
- Another avenue to look at is how do headlines fit into sentiment analysis.
- During the course of research I got to talk to an ex RTE Journalist Frances Shanahan, who was a veteran journalist specialising in agriculture. She was able to shed light on how headlines are formulated to achieve maximum effect.
- I also got to speak with Cathal O'Donnell a senior stockbroker with Cantor Fitzgerald based in St. Stephen's Green who was able to shed some light on what I should be looking for in terms of indexes etc.  
***(more detail below)***

## ***Aims***

1. Is to compile the datasets which will be crucial to the success of the overall project.
2. Is to perform some sort of ETL to the data collected to constrain the data that is necessary for us e.g. (high closing share price, timestamp etc.).
3. Is to complete the necessary documentation for the submission of the FYP.
4. Is to use Google Forms to conduct an anonymous survey for different groups of people to take. examples of groups are: students, professionals from linkedin and family members.
5. Implement some kind of machine learning into this report.
6. Get in contact with some kind of stockbroker to get insight into how sentiment operates.
7. Talk to a journalist, to see how sentiment comes into account in when they construct headlines, as headlines have to convey the whole story in a sentence.
8. The main aim of the project is to prove or disprove whether social media sentiment has an effect on share price.

## Lit review

During the process of working on my Final Year project. I had no idea what to expect when doing a Lit Review, as a result I paid a visit to the library to ask for advice and Keith told me to check out their website where I will find a helpful guide on carrying out a lit review

### Stock prediction using sentiment analysis

<http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf>

- In this case study they looked at using sentiment analysis to predict stock, this project was undertaken by Anshul Mittal and Arpit Goel both from Stanford University
- They used machine learning and sentiment analysis to see if there is a correlation between public sentiment and market price sentiment
- They used linear regression to predict future share price, based purely on the previous days price
- They used Twitter data to predict users' moods and previous days DJIA values to allow predict stock movements
- In the process of this they also implemented a stock portfolio management system based on their predicted values which they gathered previously

#### Sentiment analysis

- They went about developing their own word lists based on the POMS questionnaire Profile of Mood States
- They filtered down the number of tweets by looking for tweets that contain the words "Feel", "makes me", "I'm" & "I am" in it this allowed them to limit the sample size substantially

### Does social media affect capital markets

<https://www.forbes.com/sites/quora/2013/09/10/does-social-media-affect-capital-markets/#1ecf70a8866f>

- In this article it looks at "does social media affect capital markets" he raises some good points in relation to this and examples on companies that lost out big or gain big when social media is used correctly
- Carl Icahn is an investor and Founder, Icahn Capital Management
- Carl Icahn made an announcement of his Apple position over Twitter. Within minutes, Apple stock shot up gaining \$17 billion in market cap.
- Cleveland train crash, people started tweeting about wiped out over 500 million USD from their market cap

Why do stock prices change based on news reports?

<https://www.investopedia.com/ask/answers/155.asp>

In this article it looks at why stock prices change based on news headlines, the thing to realise is that stock price change frequently day to day this is based on supply and demand if people want to buy a particular stock, in terms of headlines negative news about a company e.g. bad earnings for a quarter and the same could be said for positive news e.g. good profits, new products etc. these normally turn into increased demand on these good performing stocks

Understanding Sentiment Analysis: What It Is & Why It's Used

<https://www.brandwatch.com/blog/understanding-sentiment-analysis>

Sentiment analysis is also known as opinion mining, it is a term used often but rarely understood by the people using it, the talk about the potential applications of sentiment analysis and that social media correlates with shifts positive or negative in the stock markets, an example of how sentiment analysis was used was by obama during the 2012 elections to gauge public opinion on how will his policy announcements and campaign messages

Data Mining

<https://searchsqlserver.techtarget.com/definition/data-mining>

Data mining is the process of sorting through large datasets to identify patterns and relationships in order to solve problems or find correlations between data to form new information, it also allows companies to predict future trends.

They then go on to talk about the four stages of data mining

- Data sources: databases
- Data exploration: sampling and transforming the dataset
- Modeling: building: a model and testing it and evaluating it
- Deploying models: using the model to influence decision making

The benefits of data mining are that it can be used to build prediction models on shoppers habits and the products they are going to buy and future products they will look for

## Meeting With Journalist Frances Shanahan

While carrying out my requirements Specification, I got to meet with a recently retired RTÉ journalist Frances Shanahan who was the Agriculture Correspondent

for RTE Radio. Frances was able to give me some insight into what to look for in headlines from the point of view of the person formulating it, she also helped me with.

### **Main Points raised by Frances Shanahan**

- Formulation of word banks and different words and what kinds of sentiment they carry.
- Provided me with an insight into how journalists formulate headlines for maximum effect.
- Also gave some points on what to what for especially sarcasm, as sarcasm is hard to pick out when conducting sentiment.

### **Meeting with a Stockbroker Cathal O'Donnell (*Cantor Fitzgerald*)**

While conducting my requirements specification I got to speak with Cathal who is a senior stockbroker with Cantor Fitzgerald based at St. Stephen's Green. Cathal has over 20 years experience in stockbroking some questions regarding the stock price index aspect of the project and his opinion on how to gauge extrinsic factors and how they affect share price.

### **Main Points raised by Cathal**

- Focus on indexes Eg. NASDAQ Composite index, Dow Jones Industrial Average, S&P500 etc.
- To use the days closing share price as the price will fluctuate wildly during the day, as a trader the closing number is what concerns him
- News headlines play a big part in his strategy when he brings forward ideas to clients, recently their analysts have started using Social media to gauge public opinion regarding

## **●.1 Technologies I planned to use**

### **MYSQL Workbench**

- This is an open-source relational database management system that I will be using to host all my datasets in a relational db which can be accessed through Tableau and or R studio, I have not come to a definite decision regarding

#### **Tweepy API/TwitterR (r Package)**

- This is used to gather tweets using the Twitter API, this api works using a Python script in which the API key is inputted and either the type of tweets is collected this can be set around different hashtags or users, for this I have chosen to focus on a single user Donald Trump as a result of my survey in which 50% of respondents said that he was the most influential person on twitter, after attempting to use it i ran into issues with the API in which if the sample size i want was set at  $n=3200$  i.e. it will return 3200 tweets from a specific user or on a hashtag. In my experiences after running the s

#### **Reddit API**

- This will be used to gather news headlines from reddit.com/r/worldnews which is a subreddit where headlines from many different sources are compiled in one place, my plan is to take these headlines with timestamps and use these as another metric within my data warehouse

#### **Google finance API**

- This will be used to collect share price information on whichever stock we have chosen for this project Cathal the stockbroker said I should focus on indexes such as the NASDAQ composite index as it gives an idea of the markets toing and froing during the weeks trading,
-

## ●.1 Technologies used

### Microsoft Excel

- Excel was used for containing and organising the initial datasets of the projects and performing rudimentary data cleansing before the dataset is used for analysis within r and tableau

### R Studio

- I was planning to use Rstudio to construct my graphs but have decided against this but in favour of Tableau due to the fact that in Tableau I can build a dashboard, but r studio will be used to construct some of the statistical tests which cannot be done in Tableau
  - Packages used
    - SentimentR
    - GGPlot
    - TwitterR
    - stringR
    - GGplot

### Tableau

- Tableau will be used for data Visualisation of the end results in an interactive way, the plan for the project is to build a dashboard using Tableau

### Google forms

- I used Google Forms to create my survey. It allows users to create surveys and interpret the results in a spreadsheet and also does visualisations of your results which you will see below in this report

# System

## ●.1 Requirements

Functional requirements

1. Pulling data from source using APIs
2. Sourcing Datasets
3. Performing Data Cleansing
4. Performing sentiment analysis of the tweets and headlines
5. Visualising results using Tableau
6. Prove or disprove the question

This section lists the functional requirements in **ranked order**. Functional requirements describe the possible effects of a software system, in other words, *what* the system must accomplish. Other kinds of requirements (such as interface requirements, performance requirements, or reliability requirements) describe *how* the system accomplishes its functional requirements. Each functional requirement should be specified in a format similar to the following:

Short, imperative sentence stating highest ranked functional requirement.

The functional requirements of the project

The functional requirements listed below are in order of priority of how (I) the User will interact with the system and its different elements in order to be able to complete this project

## Use case's

### 1. Use Case Diagram

Each requirement should be uniquely identified with a sequence number or a meaningful tag of some kind.

The Use Case Diagram provides an overview of all functional requirements.

### Requirement 1 <Creation & Populating of Datasets>

#### Description & Priority

Level1 Being Critical as the whole project depends on being able to store the datasets in a state that can be used by R studio and Tableau , without this the project cannot take place

#### Use Case

User(admin) uses a program (Excel) to creates a series of Datasets(Storage) to store three datasets

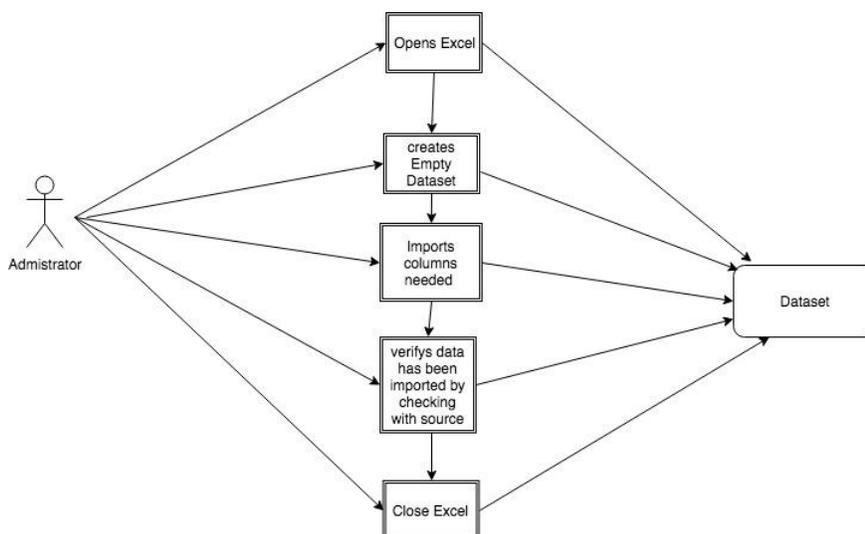
#### Scope

The scope of this use case is to use a program (excel) to create a Datasets(Storage) and so all datasets can be accessed, through a single program

#### Description

This use case describes the operation of creation of a dataset

### Use Case Diagram



## Flow Description

### Precondition

The unformed datasets must be accessible by user(Admin)

### Activation

This use case starts when an <Admin> opens program (Excel) in order to create the Dataset

### Main flow

1. The <Administrator> opens the spreadsheet application (Excel)
2. The <Administrator> creates the dataset spreadsheet
3. The <Administrator> imports the columns needed from the raw datasets
4. The <Administrator> Verifies that the Data has been inserted to the dataset by viewing the columns and cross checking them with source
5. The <Administrator> then closes the spreadsheet application (Excel)
- 6.

### Exceptional flow

E1 : <Problem with cell formatting >

1. Dates column carried over doesn't present correctly
2. The <Administor> right clicks on column and goes to format cells and checks for correct formatting
3. The use case continues at position 3 of the main flow

### Termination

When the data has been imported into the dataset

### Post condition

The data held within the Dataset and is ready to be used

## Requirement 2 <processing data>

### Description & Priority

Level 2 cleansing the datasets, to remove unrelated data columns from the datasets. this is level 2 priority as the project can still go ahead with unclesed data and ensuring date parameters are correct

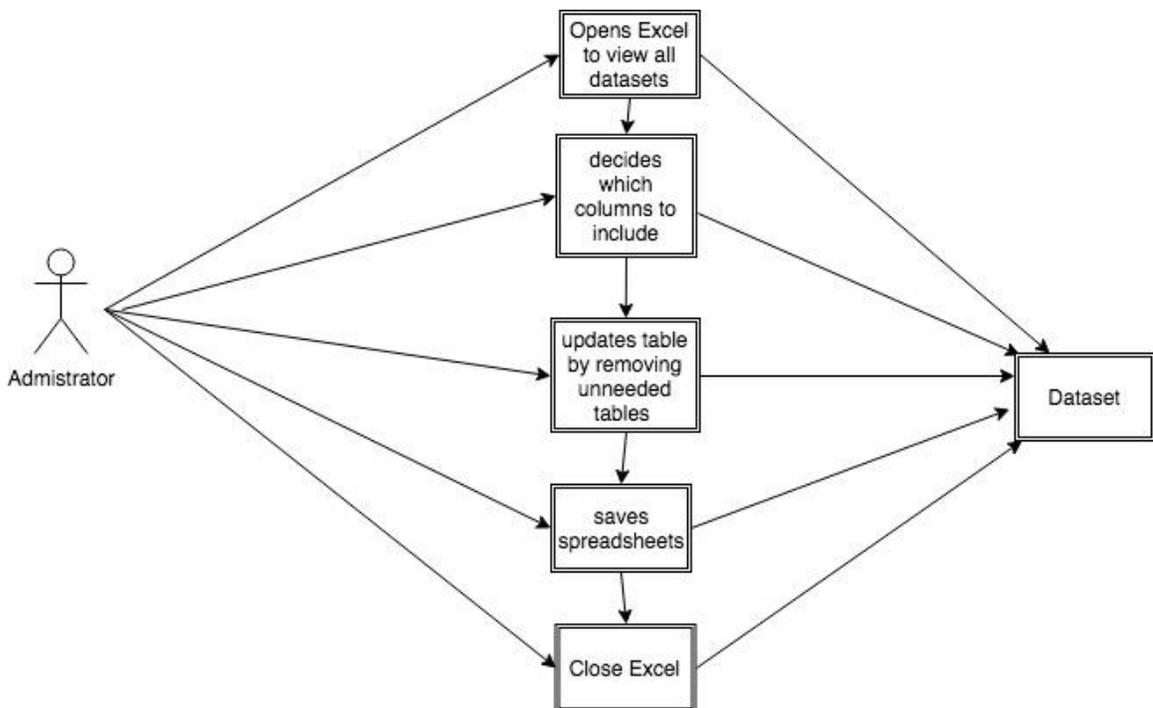
### Use Case

The administrator cleanses the data by removing unnecessary columns and formats data to lowercase

### Scope

The scope of this use case is to allow the admin to cleanse the data and constrain what types of data are to be included within the datasets

### Use Case Diagram



### Flow Description

### Precondition

The dataset to be used is in a waiting state, ready to be used

## Activation

This use case starts when an <Admin> accesses Excel

## Main flow

1. The <Admin> starts Excel to view all datasets
2. The <Admin> Decides what columns to include
3. The <Admin> then updates the table by removing the unnecessary columns
4. The <Admin> saves spreadsheets
5. The <Admin> closes Excel

## Termination

The cleansing of the dataset has been completed and the use case is terminated

## Post condition

the database goes into a waiting state waiting for user input

# Requirement 3 <Interpreting and exploring data>

## Description & Priority

Level 1 this is the main part of the project in terms of exploring the end result and as such, it is held as a level 1 priority

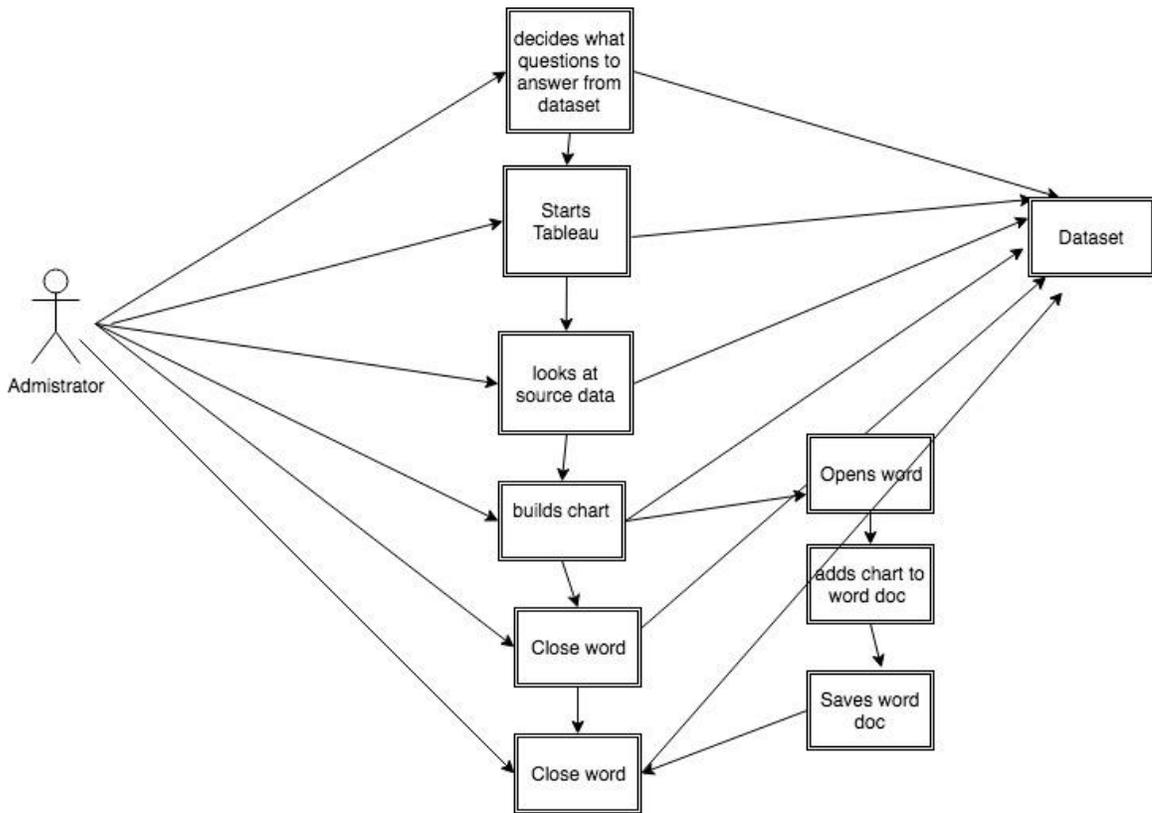
## Use Case

the administrator explores the datasets to find insights within the datasets

## Scope

The scope of this use case is to allow the admin to explore the data to find insights held within

## Use Case Diagram



## Flow Description

## Precondition

the dataset to be used is in a waiting state, ready to be used

## Activation

This use case starts when the <Admin> accesses Tableau

## Main flow

1. The <Admin> Decides on what questions want to be answered from the datasets
2. The <Admin> starts Tableau
3. The <Admin> In Tableau the Admin looks at source Data
4. The <Admin> then decides what columns to include
5. The <Admin> then builds a chart to visualise this data
  - a. The <Admin> opens Word
  - b. The <Admin> then adds this chart to the Report
  - c. The <Admin> saves word Document
6. The <Admin> Closes Word
7. The <Admin> closes Tableau

## Exceptional flow

E1 : <charts constructed in R studio>

1. The <Admin> uses R Studio to construct the Charts or perform statistical tests
2. The <Admin> collects evidence of tests and puts them into the technical report under results
3. The use case continues at position 7 of the main flow

## Termination

- When the results have been visualised and the results have been recorded and the use case is terminated

## Post condition

The system goes into a waiting state waiting for user input

# Requirement 4 <Adding Datasets to Tableau & Rstudio >

## Description & Priority

Level 1 priority as being able to add the datasets to tableau and creating the dashboard is crucial to the success of the project and as such it is held as a level 2 priority

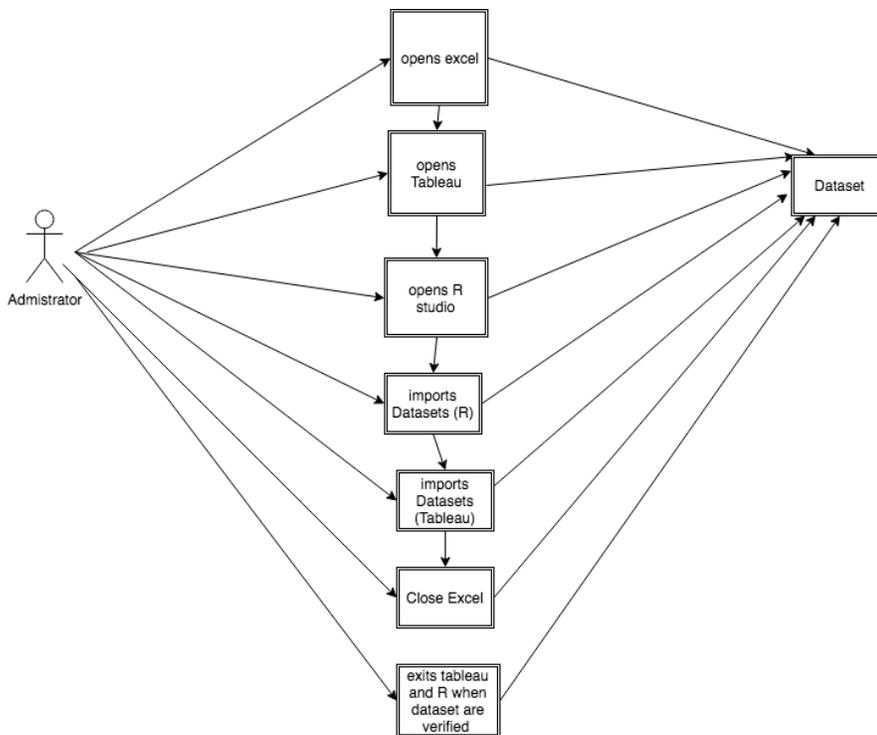
## Use Case

The administrator connects the Datasets to Tableau & R Studio So that the data can be explored

## Scope

The scope of this use case is to allow the admin to visualize the results of the data on analysis performed

## Use Case Diagram



## Flow Description

### Precondition

The dataset to be used is in a waiting state, ready to be used

## Activation

This use case starts when an <Admin> accesses Excel

## Main flow

1. The <Admin> Opens Excel
2. The <Admin> Opens Tableau
3. The <Admin> Opens R studio
4. The <Admin> imports Datasets through the add data Source option in Tableau
5. The <Admin> imports Datasets by clicking the dataset you want to use and clicking import dataset in R Studio
6. The <Admin> then Closes Excel,
7. <Admin> exits Tableau & R studio once the datasets are verified

## Exceptional flow

### E1 : <Unable to import datasets >

1. <Admin> verifies column formats
2. <Admin> then edits cell format to ensure correctness
3. the use case resumes at step 4

## Termination

When the Question has been answered by visualising certain parts of the data has been completed and the use case is terminated

## Post condition

The R studio and tableau go into a waiting state waiting for user input

## Requirement 5 <Sentiment analysis>

### Description & Priority

Level 1 this is representing the the sentiment score for the string of texts

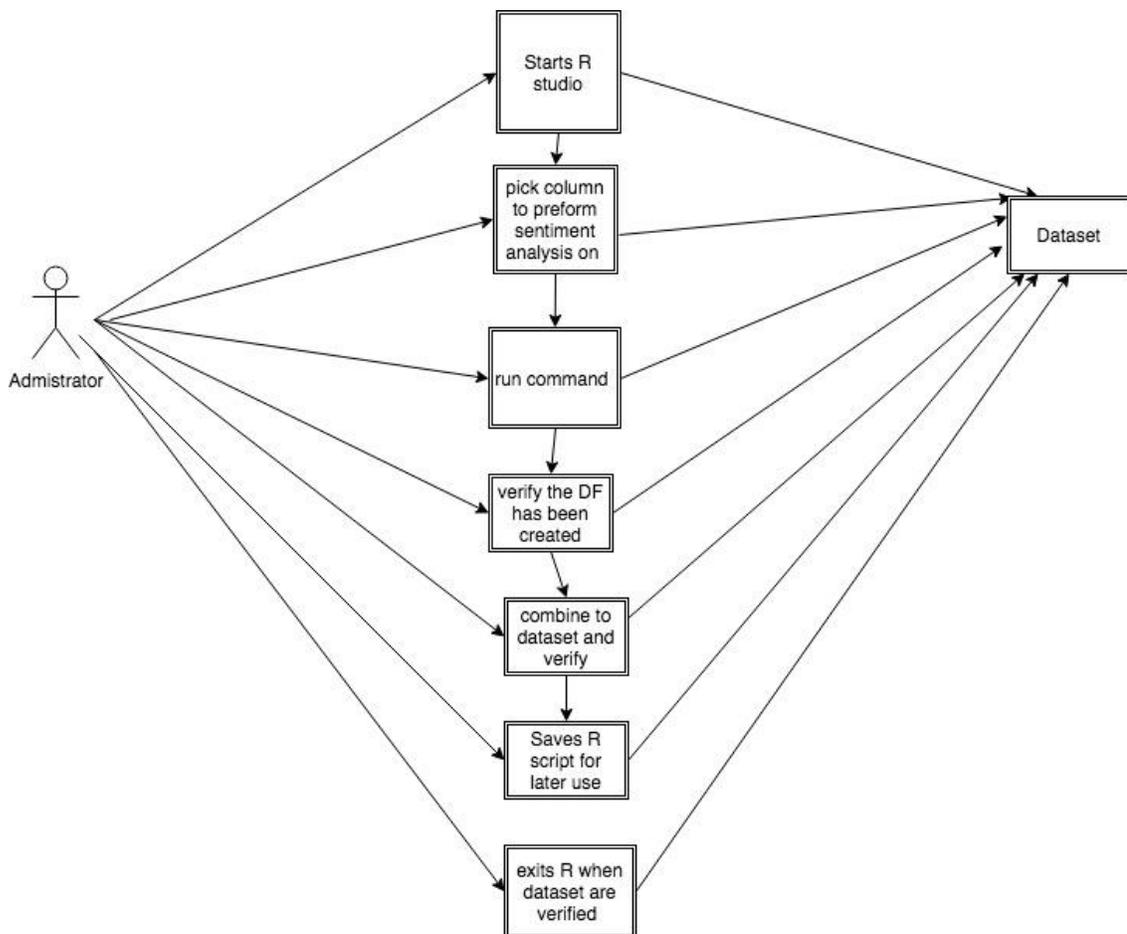
### Use Case

The administrator applies a sentiment score to the datasets

### Scope

The scope of this use case is to allow the administrator to apply a sentiment score to the datasets of tweets and headlines so that analysis can be performed

### Use Case Diagram



## Flow Description

### **Precondition**

The dataset to be used is in a waiting state, ready to be used

### **Activation**

This use case starts when an <Admin> accesses the dataset through R studio

### **Main flow**

1. The <Admin> starts R Studio
2. The <Admin> specifies which column of the dataset he wants to perform sentiment analysis on
3. The <Admin> runs the command
4. The <Admin> admin verifies the Data Frame created with scores
5. The <Admin> verifies that they have been combined correctly
6. The <Admin> saves the Rscript for later use
7. The <Admin> exits Rstudio once this is complete

### **Exceptional flow**

E1 : <Unable to perform sentiment analysis>

1. Sentiment analysis is unable to be performed
2. The <Admin> checks that the sentimentR package is installed and running
3. The use case continues at position 3 of the main flow

## Termination

When the sentiment score has been applied to

## Post condition

The database goes into a waiting state waiting for user input

List further functional requirements here, using the same structure as for Requirements 1 & 2. Most systems would have at least five main functional requirements.

## **Non-Functional Requirements**

Specifies any other particular non-functional attributes required by the system. Examples are provided below.

### **1.1.1 Performance/Response time requirement**

This project allows the user to perform analysis on the datasets provided

### **1.1.2 Availability requirement**

Datasets and the database will be available to project over the course of the project

### **1.1.3 Recover requirement**

To perform data recovery in the event of catastrophic failure of either software and or hardware I will use Github SCM and Google Drive to backup all data pertaining to the project

### **1.1.4 Robustness requirement**

Not applicable for this project

### **1.1.5 Security requirement**

Considering that we are not deploying a typical application and by which we will not be handling user information.

- Our datasets were sourced from public sources on the internet and because they are in the public domain security will not be an issue for it

### **1.1.6 Reliability requirement**

Are the datasets that I am creating reliable and correct

### **1.1.7 Maintainability requirement**

Due to the nature of the project there is no scope for maintenance due to the initial design

### **1.1.8 Portability requirement**

N/A

### **1.1.1 Extendibility requirement**

Project could be repurposed in the future but at time of writing there are no plans for expansion

### **1.1.2 Reusability requirement**

Due to the nature of project Reusability is not applicable

### **1.1.3 Resource utilization requirement**

Resources such as Hardware desktop & laptop, access to SCM and google drive, Rstudio and other programs will be needed e.g. mysql workbench

## **Data requirements**

- Having Datasets for each section Headlines, tweets and share price, laid out in a normalised table

For this section I am using 3 datasets that I plan to use for my final year Project

- Shareprice
- News Headlines
- Tweets
- These tables are laid out in a way that makes them easily usable in r studio and Tableau e.g. having the correct date format

## **User requirements**

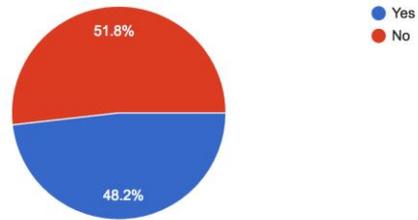
To find out the user requirements, conducted a survey and as of 8pm on the 23rd of November, 2017, 57 responses was received. The survey was distributed to my class facebook group, posted on my personal facebook, LinkedIn, Twitter, and it was sent to the company where I conducted my work placement

Q1. Do you use Twitter?

---

### Do you use Twitter?

56 responses



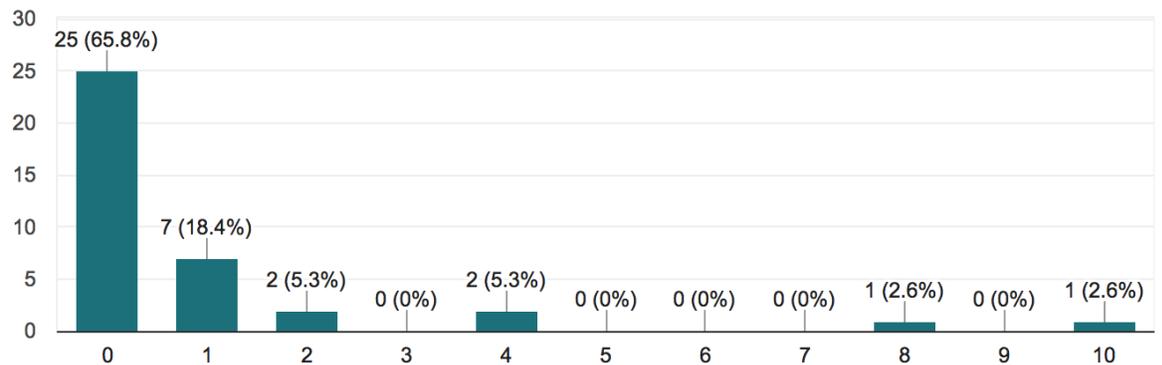
- As we can see from this chart 48% of respondents use Twitter

## Q2. What frequency of tweets do you send a day

### If yes to Q1. what frequency of tweets do you send a day



38 responses



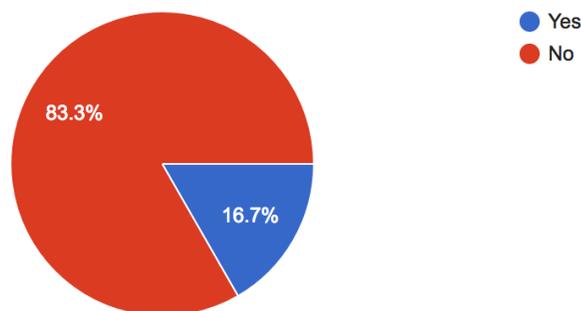
- This is a follow up from question 1 we can see that most respondents send between 1 & 4 tweets a day

## Q3. Do you Follow the stock market

### Do you Follow the stock markets?



48 responses



- most respondents don't follow the Stock Market.

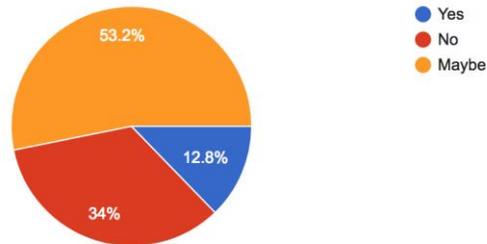
## Q4. What types of Stock do you Follow

- Tech: Snap Apple, Facebook,
- Pharma: Pfizer, Baxalta
- Index: Nasdaq
- Defence: Boeing, Airbus Lockheed Martin

Q5. Do you believe that Twitter is an accurate portrayal of current political landscape

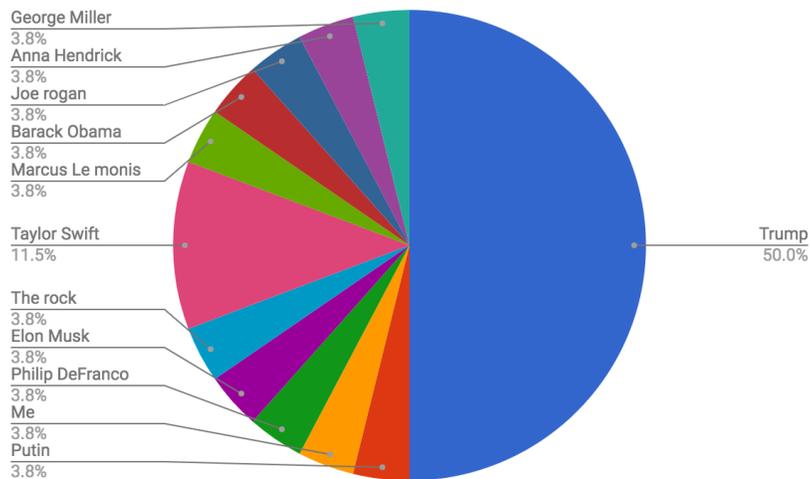
Do you believe that Twitter is an accurate portrayal of current political landscape

47 responses



- Most respondents are unsure the twitter is an accurate portrayal of the current political landscape

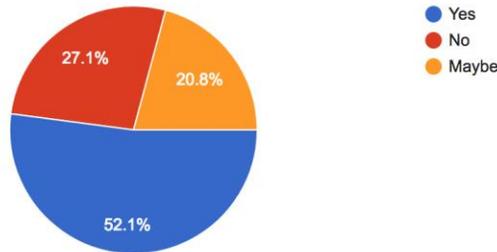
Q6. In your opinion who is the Most influential user on twitter, *Data cleansing was performed on this question to clear up some answers*



Q7. Do you believe that social media can be an extrinsic value that can affect companies share price?

Do you believe that social media can be an extrinsic value that can affect companies share price?

48 responses



- From this question we can see that most respondents believe that social media can be an extrinsic factor that affects share price

Q8. Do you believe that social media can be an extrinsic value that can affect companies share price? **IF SO WHY?**

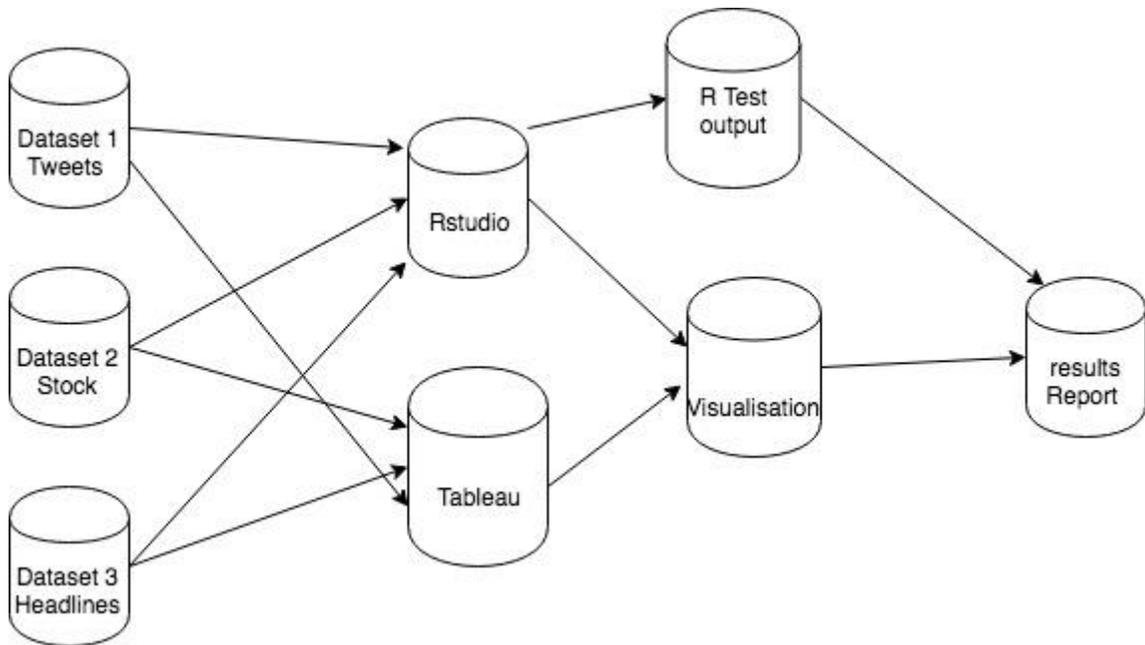
- it can be seen as as an accurate portrayal of the users feeling regarding an issue
- Look at what just happened to EA.
- Reputation at stake
- A lot of people believe what they see on social media, this could influence their decisions in buying from certain consumers, or buying/selling company stocks
- Public Image is linked to Social Image and Social Networks portray this.
- I believe that pr has large part to play in a company's success/demise, particularly now days with social media being a key player, "death by facebook" is a very real and scary thing that happens daily. It shows how society's hive mind works and how we are experiencing the modern day salem witch trials. Consider the current situation ea found themselves due to their reddit reply and the current sproi of celebrity sexual assault claims.
- Both good and bad news travel fast. Socially media speeds it up. E.g. bad reviews, celebrity endorsement etc.
- Just
- Mainly regarding crypto: Sometimes the developers may have controversial opinions which sometimes causes the prices to drop.
- Social media is advertising, advertising can lead to more investors or hype around a company. It opens the shares to the general public
- Not backed up by personal contact
- Viral social media slander campaigns are a powerful thing

- Social media is used by a huge number of people, if i follow a certain celebrity and they endorse one product over a similar product of a different brand then maybe i'm not the only one. So i do think social media can change share prices even if by a little bit
- All companies now are selling products online or through social media so depending on the amount of people who see their posts the more money they will make meaning the more followers they have the more money they will make giving them a higher share price as social media is a very good market value
- Facts are second to a good story that will "sell" in the headlines. Which can cause horrible consequences to companies.
- Anything on social media will reach everyone eventually- from your 80 year old grandad to your 7 year old nephew
- A Single Tweet can destroy a company's reputation - affecting their Share price
- It can give shareholders an insight into the company's' future behavior and decision making process

## ●.2 Design and Architecture

### System Architecture

The design architecture that I plan to use for my final year project it consists mainly of a 3 datasets which I will be



## ●.3 Implementation

### Finding Datasets

The three datasets I used were

- Tweets
  - This was sourced from <http://www.trumptwitterarchive.com/>
  - The original plan for this was to use the TwitterR package within R to use the Twitter api to pull this information, twitter only allows for 3200 tweets to be pulled at one time with the api. In reality the API will only return between 600 and 900 tweets when the script was run to interact with the twitter api
- Headlines

- I found this on Kaggle it was very difficult to find a dataset of headlines to use for a particular period of time.
- My original idea was to use Reddit's api to pull the headlines from r/worldnews but limitation in Reddit's api in terms of how far I could go back into historical data was an issue meaning I couldn't get headlines that covered my chosen
- Stock Price information
  - I compiled this dataset by search Yahoo. finance for each stock index Cathal (stockbroker) said to look at, he told me about Yahoo finance allows you to get historical data freely for most indexes
  - After compiling a few indexes I then combined them using Excel, in which I ran into a date formatting error. The solution for this was to

## **Limiting what results are needed from each dataset**

### **preprocessing of data**

- **date formatting**
  - For this I used some excel functions such as LEFT() MID() and RIGHT(). I used these to break up the existing date formats into 3 columns
- **Standardising date format**
  - Modifying the date format so that it could be interpreted by R studio and Tableau. this was done in Excel as I am proficient in using it.

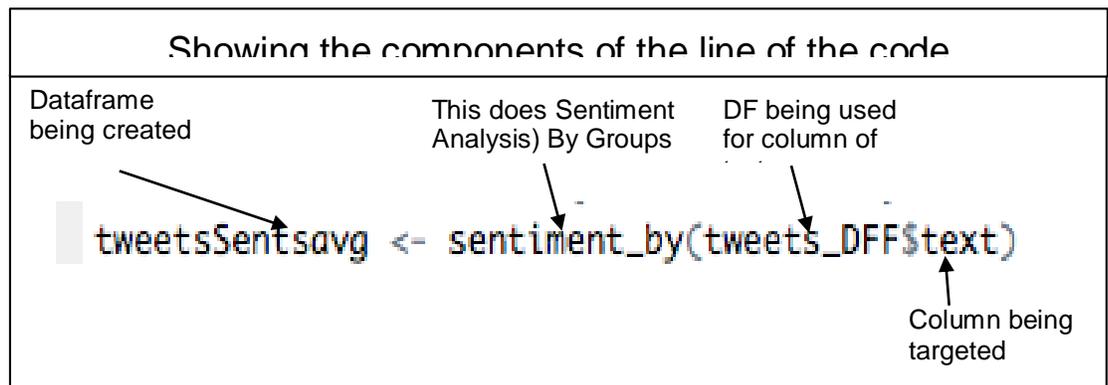
### **Filtering data**

- Data was filtered by the following parameters
  - Date
  - Changing it to lowercase
  - Removing special characters.

## Sentiment analysis

- Originally this was gonna be handled by me creating my own word banks and doing it through a Python script. upon further investigation I found there was an R package called sentimentR which I ended up using to handle my sentiment scoring

- Using SentimentR it allows for a Dataframe to be used gather a sentiment score this can be done by targeting a particular column by using the line of code below



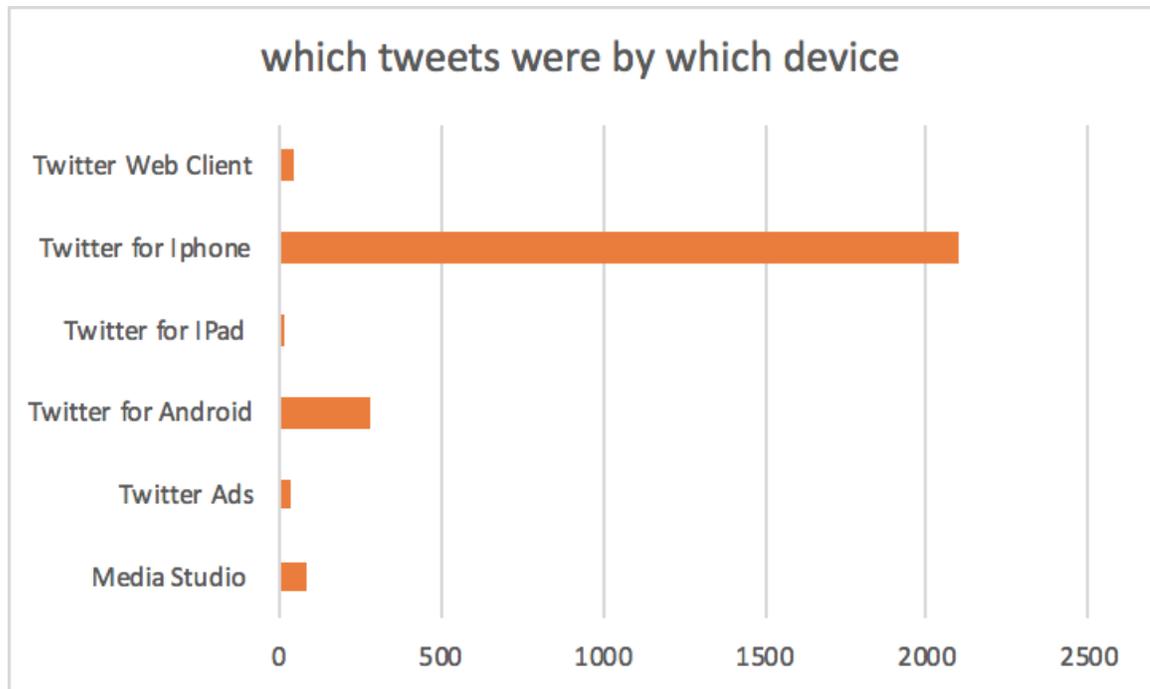
- After this line is run it produces a Dataframe containing the files below

element_id	word_count	sd	ave_sentiment
1	24	0.360652192	0.4131334939
2	22	NA	0.0000000000
3	20	0.364357520	0.4450538574
4	21	0.258825675	0.2292420316
5	15	0.258884650	0.4497257584
6	25	0.380195445	0.6672394242
7	12	0.372333805	0.4522619947
8	22	0.447721504	-0.3455143000
9	24	0.118134590	0.2915463431
10	26	0.123510924	-0.1498354123
11	22	0.331978547	-0.0182379309
12	27	0.014433757	0.0111387756
13	27	0.589300974	0.0833012851
14	29	0.152137817	0.0759549052
15	20	0.027561745	-0.3548992934
16	20	0.370779709	-0.0559701339

- Types of command to use in sentimentR
  - Sentiment gives us the Polarity Score (Sentiment Analysis)
  - Sentiment\_by gives us the Polarity Score (Sentiment Analysis) but separated into groups gives us the average

## Statistical tests

### Breakdown of tweets by device



An important breakdown is that it shows that Donald Trump switched to an iPhone on the 29th of March which would explain why the number of tweets is highest on iPhone compared to ones sent from Android.

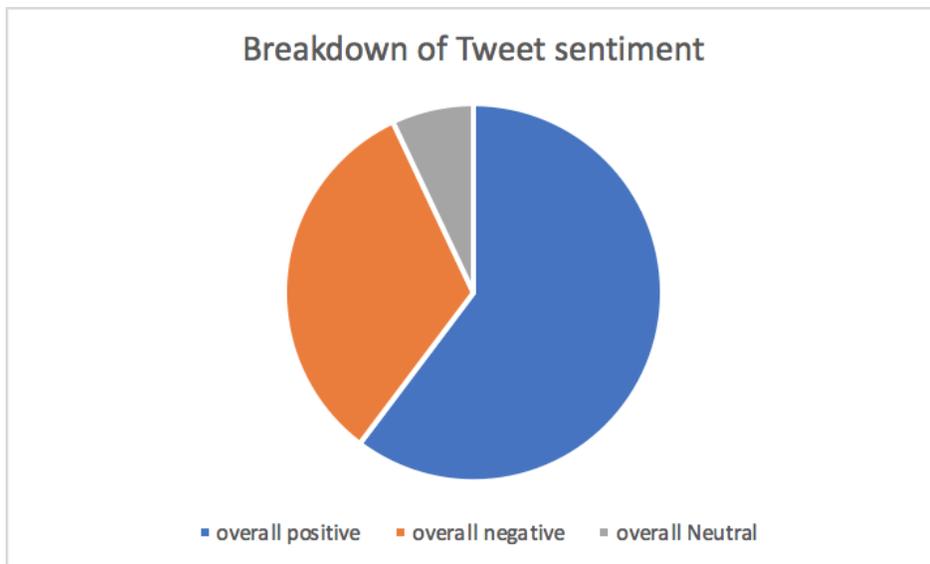
- This would also allow us to see which tweets were written by Trump and which ones were ghost written.
- As it is known that he writes all of his own tweets from his own devices e.g. iPhone until the end of March and Android for the rest of the year.

## Dan Scavino Announcing Trump's switch to iPhone



- This marks the date when Trump switched over to iPhone, this allows us to see why the number of tweets from Android was lower than iPhone

### Breakdown of tweet sentiment



## Test for normality

- Reasons for carrying out a test for normality.
  - most statistical test assume that the data is normally distributed, if the data is not-normal
  - tests for normal data are called parametric data
- P values less than 0.05 mean data are NOT normal
- P values greater than 0.05 mean data are normal
- the code below tests each value to see if the data is Normal or not-normal

### Code

```
#test for normality  
shapiro.test(DJIA)  
shapiro.test(NASDAQCOM)  
shapiro.test(SP500)  
shapiro.test(FTSE100)  
shapiro.test(SentavgTrump)  
shapiro.test(SentavgheadlineTrump)
```

### Result

```
data: DJIA  
W = 0.93352, p-value = 3.279e-09  
data: NASDAQCOM  
W = 0.97319, p-value = 0.0001128  
data: SP500  
W = 0.96594, p-value = 1.092e-05  
data: FTSE100  
W = 0.97838, p-value = 0.0007161  
data: SentavgTrump
```

W = 0.97764, p-value = 1.091e-14

data: SentavgheadlineTrump

W = 0.97764, p-value = 1.091e-14

## Descriptive Statistics

- Done in Excel its job is to describe the data
- Descriptives are split into 2 picture so that they can fit on the page

<i>headsnt_200sample</i>		<i>DJIA_200sample</i>		<i>NASDAQCOM_200sample</i>	
Mean	-0.019188001	Mean	21202.1073	Mean	6091.280593
Standard Error	0.019623265	Standard Error	56.03701127	Standard Error	21.10827293
Median	-0.033333333	Median	21054.875	Median	6145.485108
Mode	-0.033333333	Mode	#N/A	Mode	#N/A
Standard Deviation	0.27751488	Standard Deviation	792.4830134	Standard Deviation	298.5160586
Sample Variance	0.077014509	Sample Variance	628029.3265	Sample Variance	89111.83725
Kurtosis	1.744762395	Kurtosis	-0.581427983	Kurtosis	-0.917699338
Skewness	-0.351271007	Skewness	0.174669993	Skewness	-0.270139522
Range	2.007918077	Range	3265.03906	Range	1194.919922
Minimum	-1.025039651	Minimum	19732.40039	Minimum	5429.080078
Maximum	0.982878426	Maximum	22997.43945	Maximum	6624
Sum	-3.837600204	Sum	4240421.459	Sum	1218256.119
Count	200	Count	200	Count	200
Confidence Level(95.0%)	0.038696227	Confidence Level(95.0%)	110.5025511	Confidence Level(95.0%)	41.62459695

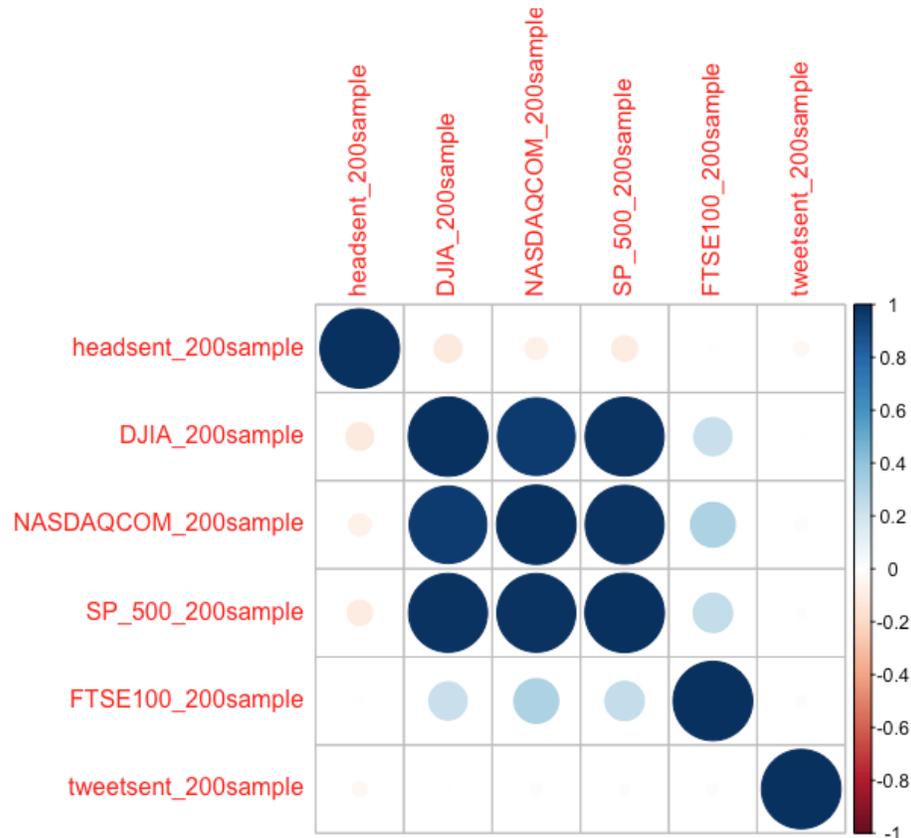
<i>SP_500_200sample</i>		<i>FTSE100_200sample</i>		<i>tweetsnt_200sample</i>	
Mean	2406.43125	Mean	7397.68	Mean	0.051435602
Standard Error	5.273016272	Standard Error	6.660213225	Standard Error	0.019838074
Median	2410.775	Median	7389.7	Median	0.050279785
Mode	2268.9	Mode	7547.6	Mode	0
Standard Deviation	74.57171126	Standard Deviation	94.18963871	Standard Deviation	0.280552731
Sample Variance	5560.940121	Sample Variance	8871.68804	Sample Variance	0.078709835
Kurtosis	-0.598256303	Kurtosis	-0.005780028	Kurtosis	1.53423256
Skewness	-0.09387525	Skewness	-0.368054256	Skewness	0.136430512
Range	301.53	Range	447.9	Range	1.995522579
Minimum	2257.83	Minimum	7114.4	Minimum	-0.767522579
Maximum	2559.36	Maximum	7562.3	Maximum	1.228
Sum	481286.25	Sum	1479536	Sum	10.28712039
Count	200	Count	200	Count	200
Confidence Level(95.0%)	10.39815895	Confidence Level(95.0%)	13.13365105	Confidence Level(95.0%)	0.03911982

## Correlation matrix

- why use a correlation matrix

- a correlation matrix is used to see the correlations between different types of data in this case it will be between -1 to 0 to +1 -1 and +1 being the strongest correlations
- the sample used below is of the first 200 rows

Correlogram plot of news & Tweet average sentiment vs share price



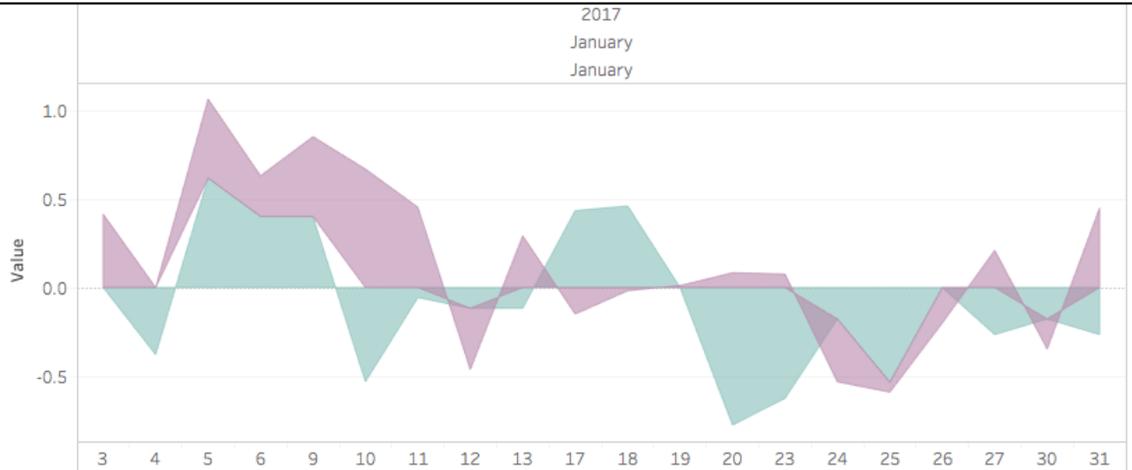
```

> corMat
      headsent_200sample DJIA_200sample NASDAQCOM_200sample SP_500_200sample FTSE100_200sample tweetsent_200sample
headsent_200sample  1.000000000 -0.119335540 -0.07512944 -0.1000264 -0.008391981 -0.032722446
DJIA_200sample     -0.119335540  1.000000000  0.95296337  0.9848613  0.229678016  0.001909311
NASDAQCOM_200sample -0.075129436  0.952963368  1.00000000  0.9828871  0.315302126 -0.019718608
SP_500_200sample   -0.100026360  0.984861334  0.98288709  1.00000000  0.242995866 -0.010813995
FTSE100_200sample  -0.008391981  0.229678016  0.31530213  0.2429959  1.000000000  0.015807897
tweetsent_200sample -0.032722446  0.001909311 -0.01971861 -0.0108140  0.015807897  1.000000000
> corrrplot(corMat)

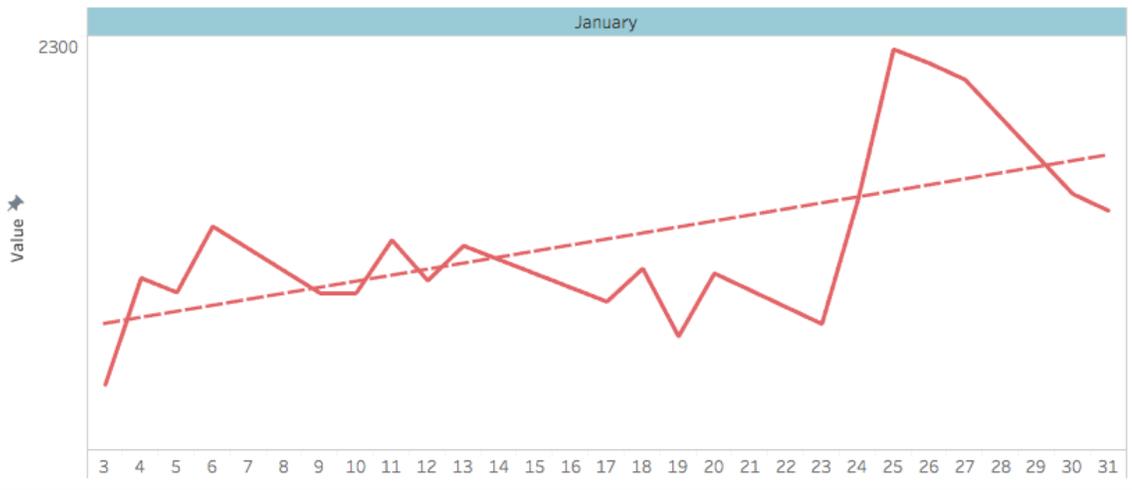
```

correlation	headsent_200sample	DJIA_200sample	NASDAQCOM_200sample	SP_500_200sample	FTSE100_200sample	tweetsent_200sample
headsent_200sample	1					
DJIA_200sample	-0.11933554	1				
NASDAQCOM_200sample	-0.075129436	0.952963368	1			
SP_500_200sample	-0.10002636	0.984861334	0.982887091	1		
FTSE100_200sample	-0.008391981	0.229678016	0.315302126	0.242995866	1	
tweetsent_200sample	-0.032722446	0.001909311	-0.019718608	-0.010813995	0.015807897	1

### Graph of Headline & Tweet average Sentiment for January



### <S&P 500 Shareprice with trend (JAN)>



#### **●.4 Graphical User Interface (GUI) Layout**

Initially I hadn't planned to implement a Graphical User Interface (GUI), I had planned to compile charts in R, but upon research I have decided on using Tableau which is a Data Visualisation program. Tableau Will be used to create a create a dashboard, so that I can visualise my results in an interactive way and will compliment some of the other tests in R studio.

#### **●.5 Application Programming Interfaces (API)**

- I planned on using the api's below but in reality they wouldn't return the right amount of data that i needed e.g. date parameters, for example the twitter api would only return between 600 and 840 tweets when i ran the python script
  - Tweepy is a python library which is to be used for accessing the twitter API so that we can pull down the Tweets saved as a .csv file
  - Reddit API: this will be used for pulling the headlines for reddit.com/r/worldnews, these headlines will be used as another metric in my overall analysis
  - Finanace.Google.com this will be used for gathering stock price index information, through a Python script

#### **●.6 Testing**

##### ***User Testing***

- Due to the nature of the project it was hard to figure out some sort of testing, as a result I had a chat with Eugene O'Loughlin who was my

Business Data analysis lecturer on what I should do. He said to do some sort of user testing which is what I ended up doing.

- I got 3 users to run my code and run the commands to see if they got the same results, I did this to ensure that my code works as expected
- On the user testing sheet below the steps of the test are
- I used the rough template of a user test i did when i was on work placement

### ●.1 ***User testing table***

- User testing was performed by giving the below table and a copy of my code and dataset to 3 users for them to go through to see if they can replicate the results
- results will be included in the appendices
- All tests carried out by participants passed as per the procedure

ID	TASK	Steps	Pass/Fail Yes/no	Issues
1	Create datasets from .csv files	<ol style="list-style-type: none"> <li>1. open R-studio</li> <li>2. Set working directory as fypcode</li> <li>3. right click on each .csv and hit import dataset</li> <li>4. ensure the data types are correct</li> <li>5. click import</li> <li>6. repeat for each dataset</li> <li>7. verify DF have been created by looking at</li> </ol>		
2	Run sentimentR package	<ol style="list-style-type: none"> <li>1. open R studio</li> <li>2. Install the package using <code>install.packages("sentimentr")</code></li> <li>3. Ensure <code>library(sentimentr)</code> is running</li> <li>4. Highlight and run both of these</li> <li>5. Use this line of code to get sentiment score for tweets</li> <li>6. <code>tweetsSentsavg &lt;- sentiment_by(tweets_DF\$text)</code></li> <li>7. A DF called <code>tweetsSentsavg</code> will be created and should be listed in Data environment</li> </ol>		
3	Combine sentiment Dataframe with tweets & headline DF	<ol style="list-style-type: none"> <li>1. open R studio</li> <li>2. enter command <ol style="list-style-type: none"> <li>a. <code>b &lt;- tweets_april[,c(1)]</code></li> <li>b. <code>c &lt;- tweets_april[,c(2)]</code></li> <li>c. <code>d &lt;- tweets_april[,c(3)]</code></li> <li>d. <code>e &lt;- tweets_april[,c(4)]</code></li> <li>e. <code>f &lt;- tweets_april[,c(5)]</code></li> <li>f. <code>g &lt;- tweets_april[,c(6)]</code></li> <li>g. <code>h &lt;- tweets_april[,c(7)]</code></li> <li>h. <code>i &lt;- tweetsSentsavg[,c(3)]</code></li> <li>i. <code>J &lt;- tweetsSentsavg[,c(4)]</code></li> <li>j. <code>k &lt;- tweetsSentsavg[,c(3)]</code></li> <li>k. <code>tweets_DF_sentAvg = data.frame(b, b, c, d, e, f, g, h, i, J, k)</code></li> </ol> </li> <li>3. Verify the DF has been created by typing <code>tweets_DF_sentAvg</code> and highlight it and hit run (top right of the window)</li> <li>4.</li> </ol>		

4	Convert to value	<ol style="list-style-type: none"> <li>1. Open rStudio</li> <li>2. Run these commands <ul style="list-style-type: none"> <li>○ SentavgheadlineTrump &lt;- headlineDSTrump\$ave_sentiment</li> <li>○ sentavgtweet &lt;- tweetsDS\$ave_sentiment</li> <li>○ DJIA&lt;- stockprices1\$DJIA</li> <li>○ NASDAQCOM&lt;- stockprices1\$NASDAQCOM</li> <li>○ SP500 &lt;- stockprices1\$SP_500</li> <li>○ FTSE100&lt;- stockprices1\$FTSE100</li> </ul> </li> <li>3. verify the value have been created by typing the name of each value in the script and run them, there values should be listed in the console</li> </ol>		
5	Shapiro Wilk test	<ol style="list-style-type: none"> <li>1. Open Rstudio</li> <li>2. Run command</li> <li>3. shapiro.test(DJIA) (<b>VALUE WHICH YOU WANT TO TEST</b>)</li> <li>4. Record result</li> <li>5. Result should be: <math>W = 0.93352</math>, <math>p\text{-value} = 3.279e-09</math></li> </ol>		
6	Wilcoxon test	<ol style="list-style-type: none"> <li>1. Open Rstudio</li> <li>2. Run Command</li> <li>3. wilcox.test(SentavgTrump, DJIA)</li> <li>4. wilcox.test(DJIA, SentavgTrump)</li> <li>5. Record Result:</li> <li>6. Result should be <ul style="list-style-type: none"> <li>○ <math>W = 0</math>, <math>p\text{-value} &lt; 2.2e-16</math></li> <li>○ <math>W = 383780</math>, <math>p\text{-value} &lt; 2.2e-16</math></li> </ul> </li> </ol>		
	Tableau	<ol style="list-style-type: none"> <li>1. Open tableau</li> <li>2. Import dataset into tableau by using new data source</li> <li>3. Ensure that the correct date format is used</li> <li>4. To test drag "Date Manual" into column</li> <li>5. Drag DJIA into rows</li> <li>6. A bar chart should be formed</li> <li>7. Verify that it is correct by click on the[+] icon on year until the day parameter shows up then remove the others</li> <li>8. Right click on DJIA in rows and change the</li> </ol>		

		measure to AVG		

Name: \_\_\_\_\_

Signature: \_\_\_\_\_

Date: \_\_\_\_\_

## ●.1 *Evaluation*

- We ended up not using the Python scripts to gather the datasets, as we ran into issues getting the correct data for a particular date period.
- In relation to the stock Dataset all providers of financial information hold their api's behind paywalls so this eliminated this for me
- It allowed me to choose Donald Trump as my tweets Dataset as % people said he was the most influential person on twitter
- My use cases had to be completely changed due to not using a DB/DW to hold my info and not using API's the intended way

## ● **Results**

As a result of carrying out this research I can see that there is a low to medium correlation between share price primarily FTSE100 and Tweet Sentiment, and a low correlation with headlines this shows us that tweets could be used a medium to predict anticipate share price peaks and dips, careful consideration needs to be taken in account regarding Sarcasm as sentiment analysis cant pick up on it and introduce some false negatives

## ● Conclusions

### Advantages

- Scalable idea can be implemented it with different metrics to see if sentiment is an extrinsic factor within different types of data i.e. sales data for a company
- Every company that operates in the public eye can use sentiment analysis to gauge how happy their customers are using their product or service. an example would be irish rail checking for tweets where people use negative words

### Disadvantages

- Comparing sentiments of different datasets is near impossible as it is counting the number of positive words present
- Sentiment analysis cannot interpret sarcasm

### Opportunities

- Project could yield interesting results if additional datasets were to be used.

## Further development or research

After completing all aspects of this Final Year project

With more resources, where could the results of this project lead to?

### Lit review

We doing the lit review was one of the more difficult sections of the fyp due to leaving it to the final weeks, this should of been done and dusted early in the first semester in all honesty, I feel that it sets us the progress once it is done, also if you can find a similar project in the same area you are trying to address can be a great help

### Survey

- I felt I achieved a lot in terms of the survey. I learned the hard way is to check the wording of the questions again and again (measure twice and cut once) applies here. ensure the questions can be understood by a 5 year old, don't load them up with technical jargon
- If i had more time i would of done a more in depth survey rather than the one I did with ten questions

### Sentiment analysis

- In hindsight it would of been a lot better if i wrote the python script to handle sentiment analysis but seen as R had a package i felt it was a waste to make my own i.e. not reinventing the wheel, I would of like to to analysis on the whole dataset at the end rather the first 200 of each dataset that I chose in the end but I felt that it was too confusing to do this considering my knowledge and skills with r and data analytics

### Overall

- Maybe I would of picked an easier topic one with a lot less variables to be considered
- Add datasets to a DB of sorts probably a mysql db then connect that DB to R studio and Tableau
- Automation of the data gathering process by running a python script in a loop on a virtual microsoft Server, to continuously update the DB with fresh info and apply sentiment score to that data
- Figure out how to work the Sentiment analysis to update either the source data or the DB itself
- Deploying my Tableau Dashboard to a web page
- Perform more in depth Data Cleansing

## ● References

- Hashimoto, S. (2018). *7 Benefits of Sentiment Analysis You Can't Overlook*. [online] Blog.insightsatlas.com. Available at: <https://blog.insightsatlas.com/7-benefits-of-sentiment-analysis-you-cant-overlook> [Accessed 11 May 2018].
- Cs229.stanford.edu. (2018). [online] Available at: <http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf> [Accessed 11 May 2018].
- Forbes.com. (2018). *Forbes Welcome*. [online] Available at: <https://www.forbes.com/sites/quora/2013/09/10/does-social-media-affect-capital-markets/#1ecf70a8866f> [Accessed 11 May 2018].
- Beers, B. (2018). *Why do stock prices change based on news reports?*. [online] Investopedia. Available at: <https://www.investopedia.com/ask/answers/155.asp> [Accessed 11 May 2018].

- Brandwatch. (2018). *Sentiment Analysis: How Does It Work? Why Should We Use It?*. [online] Available at: <https://www.brandwatch.com/blog/understanding-sentiment-analysis> [Accessed 11 May 2018].

- **Appendix**

- .1 ***Project Proposal***

- .2 ***Monthly reports***

- .3 ***User testing sheets from users***

## 1. Objectives

The objective of this project is to prove or disprove that donald trump voice on social media focusing on Twitter is an extrinsic factor, taking into account headlines are factors that affects share price of Defence/Aerospace and or Fortune 500 companies

***The project can be broken down into three main parts***

<b>Data Gathering</b>	<b>Sentiment</b>	<b>Results/Visualisation</b>
Gathering Data from different publicly available sources and perform ETL	using sentiment analysis to judge the outcome of strings of text	Using sentiment to prove/ Disprove the Objective

This implementation will be built iteratively in different stages

1. Gather stock data using python scripts
2. Gather POTUS tweets using the twitter API
3. Adding another metric to use eg. news headlines all linked to timestamp
4. Create word banks for positive/ negative
5. Use python or R to perform sentiment analysis

6. find coloration between different datasets
7. Visualise results
8. Prove or disprove results

## **2. Background research**

- I've been thinking a lot about doing an Analytical type project for my final year, this was leveraged by work placement coordinator during my time with TQS Integration as it is something that would stand out in an interview, as opposed to the normal way of developing an App etc..
- the precursor to this is my avid interest in the stock market especially how it works and is able to influence the world economy
- i've always loved how twitter is always in the news and the way it is used to gauge of opinion, even if twitter's usage is still quiet niche in terms of total active users and the demographics of people who use twitter
- I came up with the project idea around the time of when North Korea were threatening the Us territory of Guam as i noticed during that week Donald Trump's twitter was being quite forward and convincing the world that we were on the brink of nuclear war, i also noticed that during this period that the share prices of certain aerospace companies share prices increased dramatically during this period as well as other types of industries that supply the aerospace industry
- another avenue to look at is how do headlines fit into sentiment analysis
-

### **3. Technical Approach**

**Brief description of the approach to be followed (Max. 1 Page), Research, literature review, requirements capture, implementation etc...**

this projects consists of the following sections

- Mining tweets
  - i will have to investigate a way of mining tweets either from a singular person or from
  - i have looking to using twitter's api for this to work you need an api key which can be got here <https://apps.twitter.com/app/new>
- Mining Share price data
  - my original plan was to use google.finance
  - <https://chrisconlan.com/download-daily-data-every-sp-500-stock-r/>
- Mining News headlines
  - for this I have planned to use reddit for gathering headlines from a subreddit called r/worldnews which would allow me to gather these headlines and store them as a string
- generating word banks
  - my plan is to create word banks for different types of sentiment e.g. positive words and bad but i could also break these down into emotions e.g. fear anger happiness etc.
- performing actual sentiment
  - creating the codebase to perform this function
- Visualisation of results
  - i'll need to determine how best to represent my results
  - do i either
    - develop a dashboard
    - visualise through charts etc. and explain my results

to organise my thoughts and to formulate a process i have used trello, which allows me to organise thought pertaining to the project

#### **4. Special resources required**

- Python
- Rstudio
- Server hosting to have the python scripts running constantly. either ubuntu or MS Server 2012 r2
  - <https://cyclades.oceanos-global.grnet.gr/ui/>

#### **1. Technical Details**

Implementation language and principal libraries

- R
  - I plan to use R To visualise my results
- Python
  - Python will be used to run the scripts that interact with the different api's to pull information into a .csv file. from which these can be put through some form of etl and them inserted into the DB
- Twitter API
  - I plan to use this API to pull Tweets into a .csv so that they can be added to the DB
- Reddit API

## **2. Evaluation**

Describe how you will evaluate the system with real technical data using system tests, integration tests etc. In addition, where possible describe how you will evaluate the system with an end user.

I plan to evaluate the project based on the following criteria:

Data Collected from sources	can Sentiment be automated
Data undergoes some form of etl	can new tweets and such be added to the dataset automatically
results are visualised correctly	
how deep can the datasets be sorted by	Can the collection of data be automated

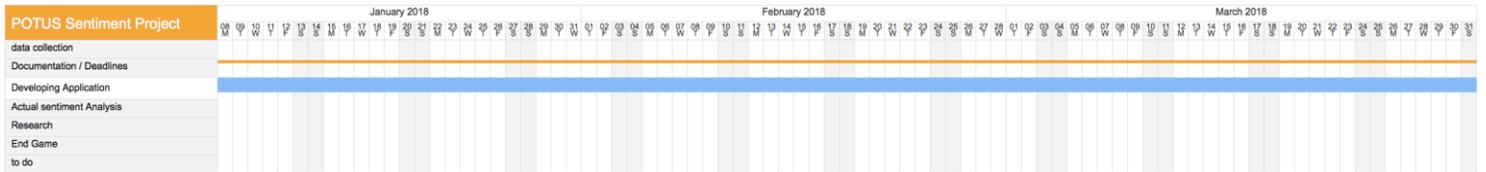
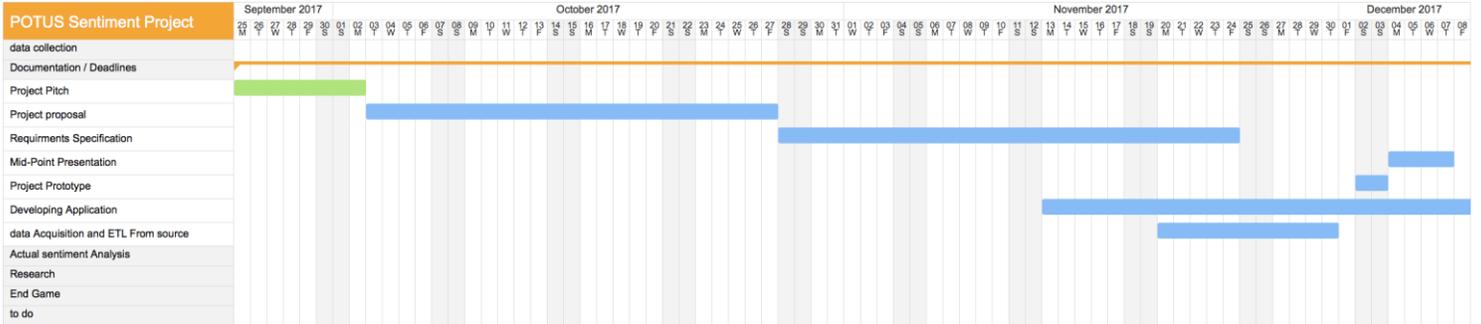
## **5. Technical Details**

Implementation language and principal libraries

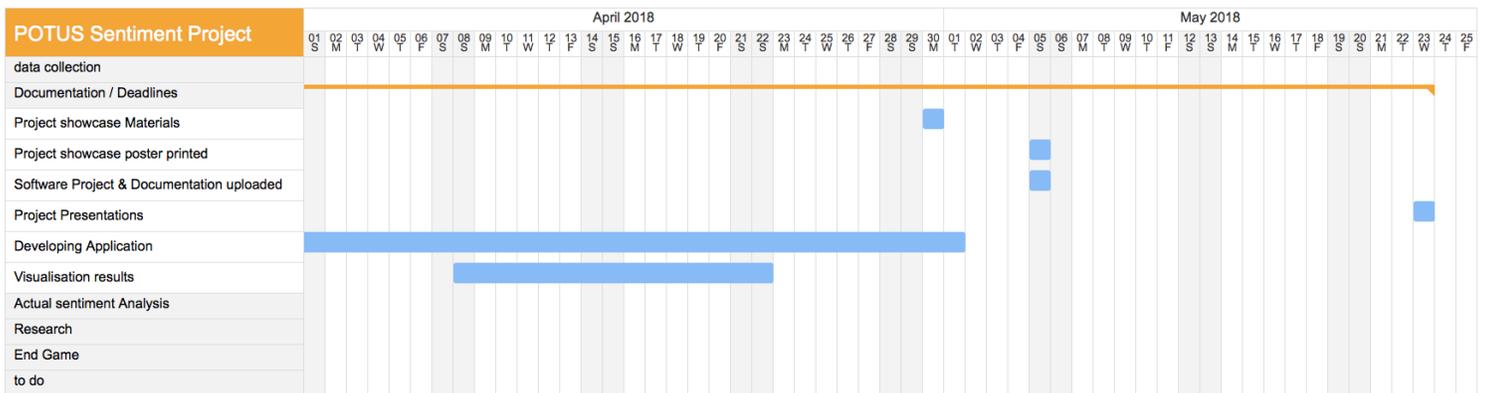
- R
  - I plan to use R To visualise my results
- Python
  - Python will be used to run the scripts that interact with the different api's to pull information into a .csv file. from which these can be put through some form of etl and them inserted into the DB
- Twitter API
  - I plan to use this API to pull Tweets into a .csv so that they can be added to the DB
- Reddit API

## ●.4 Project Plan

### Semester 1



### Semester 2 Jan to March



### Semester 2 April to June

## ●.5 *Monthly Journals*

### Monthly reports

SEP

Student name: Emlyn Farrell

Programme (e.g., BSc in Business Information Systems):

Month: September

My Achievements

- was able to brainstorm ideas and come up with a viable proposal to be presented to the lecturers during the pitch
- plot out the Project idea into main points & also outline the technologies that are need to accomplish this
- being able to correctly convey my ideas to the lecturer during pitch

My Reflection

- However, I was not successful in my time management as i felt i didn't work as efficiently as i would of thought, but overall i was happy in the way i worked, especially since

Intended Changes

- i need to nail the scope the of the project idea before i fill out the proposal so that there is a solid basis to build on
- implement a trello board for plotting out functionality and other ideas pertaining to the project

OCT

Student name: Emlyn Farrell

Programme (e.g., BSc in Business information systems):

Month: oct/nov

You don't have to follow the suggested format. These sub-headings and questions below may help you to get the most out of this journal, but you are free to modify as you see fit. Through this journal you demonstrate that you are engaged with the process and that you can identify what you need to do or change to progress and succeed in this project.

Upload one journal every month. Expected word count 300 words (of you own words).

My Achievements

This month, I was able to move forward steadily with project as a whole

I got to meet to meet stockbroker cathal O'Donnell from the firm Cantor Fitzgerald. He was able to give me great insight into what they look for in relation to Extrinsic factors

used the example of EA and the Battlefront 2 Fallout and generally explained a few things in relation to stocks, overall was impressed with what I was doing  
I got to meet with Retired Radio journalist Francis Shanahan who used to work for Rte. Radio as an agriculture correspondent, she was able to help me come up with different groups of words to conduct sentiment with. She also gave me an insight into the keywords that go into creating headlines  
I worked out the technical aspects of the project  
I met with Lisa my project supervisor and had a talk over the project ideas and ways to further it

#### My Reflection

I felt, it worked well to meet with SME's in relation to the perspective fields  
However, I was not successful in Time management. this is something to work on in future considering how tight the schedule is taking into account the other modules

#### Intended Changes

Next month, I will try to be more time conscious in relation to deadlines and hitting all deliverables on time  
I realised that I need to conduct more research about twitter impact on markets in a non-direct way and in more detail

#### Supervisor Meetings

Date of Meeting: 31/10/17

Items discussed: moving forward to requirements specification, testing, Requirements specification

Action Items: conduct a survey, think up of some testing protocols

## NOV

#### **My Achievements**

This month, I was able to move forward steadily with project as a whole, i started on working towards the midpoint presentation which involved finalising documentation, and

finalizing some sort of prototype for the midpoint, for the prototype i compiled 2 datasets , share price and tweets, which i was going to use for my project I met with Lisa my project supervisor and had a talk over the project ideas and ways to further it including a project plan and what she expects to see for midpoint presentation

### **My Reflection**

Time management is something that i need to work on considering how tight the schedule is taking into account the other modules

### **Intended Changes**

Next month, I will try to be more time conscious in relation to deadlines and hitting all deliverables on time I realised that I need to conduct more research about twitter impact on markets in a non-direct way and in more detail

JAN

Month: january

## **My Achievements**

- I was able to brainstorm ideas and come up with a viable proposal to be presented to the lecturers during the pitch
- plot out the Project idea into main points & also outline the technologies that are need to accomplish this
- being able to correctly convey my ideas to the lecturer during pitch

## **My Reflection**

- However, I was not successful in my time management as i felt i didn't work as efficiently as i would of thought, but overall i was happy in the way i worked, especially since
- 

## **Intended Changes**

- i need to nail the scope the of the project idea before i fill out the proposal so that there is a solid basis to build on
- implement a trello board for plotting out functionality and other ideas pertaining to the project

MAR

### **My Achievements**

got sentiment to work with a sample data there was some teething issues with this due to the way i had my dataset laid out and also getting to know how dataframes and dataset work within the r environment

i was able to visualise some result in r as well as tableau, in tableau i have graphed the sentiment of both both datasets together based on date to see if there is a correlation between the 2 datasets by taking an average for each day a limit of this i discovered is that the length of the text will determine whether it will be more positive than the other dataset,

### **My Reflection**

I got a lot more than i did last month in terms of fyp and the project is starting to come together i feel, i feel a lot more comfortable about it now

### **Intended Changes**

finalise methods and update the documentation and use whole dataset to process data through through

### **Supervisor Meetings**

we had a meeting but forgot what was discussed, it was positive

#### **●.6 Other Material Used**

Any other reference material used in the project for example evaluation surveys etc.