# Twitter Sentiment Analysis

# #Brexit

Data Mining

M.Sc. Data Analytics – Group D

Poonam Dhoot - 10399137

# Acknowledgement

I would like to acknowledge and thank the following people for their support and motivation during this project.

I would like to express my sincere gratitude and appreciation towards my Professor **Terri Hoare,** without whose guidance and mentoring I would not have been able to complete this project. Her words of constant encouragement and meticulous reviews have helped me improve the work and its quality in many aspects. Short and succinct meetings with her has played a key role in this project and has helped me stay on top of the schedule.

Moreover, I would also like to thank my project partner **Mr. Sunmeet Thapar** for his idea, efforts, constant support and motivation without which the project would not have been completed successfully.

Finally, I would like to thank my family for their full-fledged support, without which pursuing this degree wouldn't have been possible.

# Table of Contents

# Introduction

Twitter Sentiment Analysis is a project based on big data analytics. This project will help us to analyze sentiment based on the tweets on a particular topic. Sentiment Analysis is the process of '*computationally*' determining whether a piece of writing is positive, negative or neutral. It is also called as Opinion Mining. It is used to understand how people feel about the topic based on the tweets on that specific topic.

The topic we have chosen here to analyze the sentiment is Brexit.

What is Brexit?

Brexit is short for "British exit" - and is the word people use to refer to the United Kingdom's decision to leave the European Union (EU). The EU is a political and economic union of 28 countries which trade with each other and allow citizens to move easily between the countries to live and work.

Why Brexit?

A public vote - called **a referendum** - was held on Thursday 23 June 2016 when voters were asked just one question - whether the UK should leave or remain in the European Union.
The Leave side won by nearly 52% to 48% - 17.4m votes to 16.1m - but the exit didn't happen straight away. It was due to take place on 29 March 2019.

Now that Brexit has passed, using Twitter Sentiment Analysis we can find out the overall sentiment on Brexit. Negative being people are not happy with the Brexit. Positive being they are in favor of Brexit. Neutral being they are indifferent with it.

## Business Understanding

Social media is used by people and politicians to prove their point and as a result there were comments, tweets, and posts in support and against of Brexit. Following this example, here we are going to familiarize ourselves with Aylien (library with auto sentiment classification) by determining the sentiments of Brexit tweets as positive, negative and neutral.

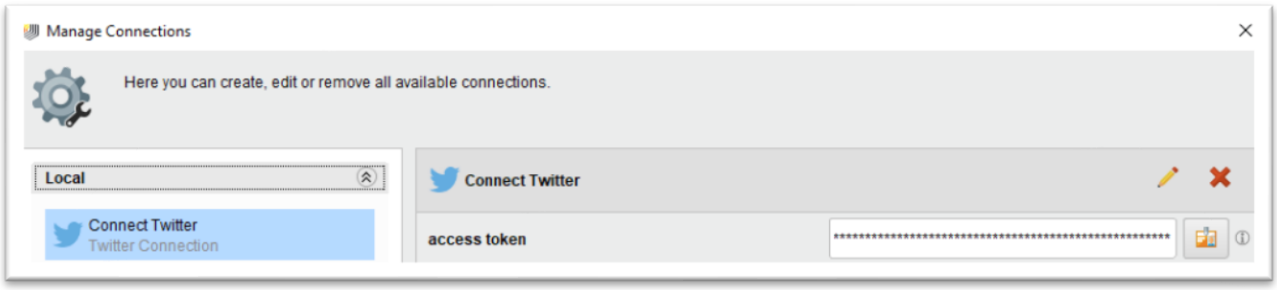
“Text Analysis by AYLIEN” is an Extension made up of different Operators that allows us to analyze and make sense of textual data from within RapidMiner. The different Operators contained in “Text Analysis by AYLIEN” include the following:

- Sentiment Analysis
- Entity Extraction
- Language Detection
- Hashtag Suggestion
- Related Phrases

# Data Understanding

Live data was extracted from Twitter API using RapidMiner's in-built operator group – Data Access which provides Search Twitter operator.

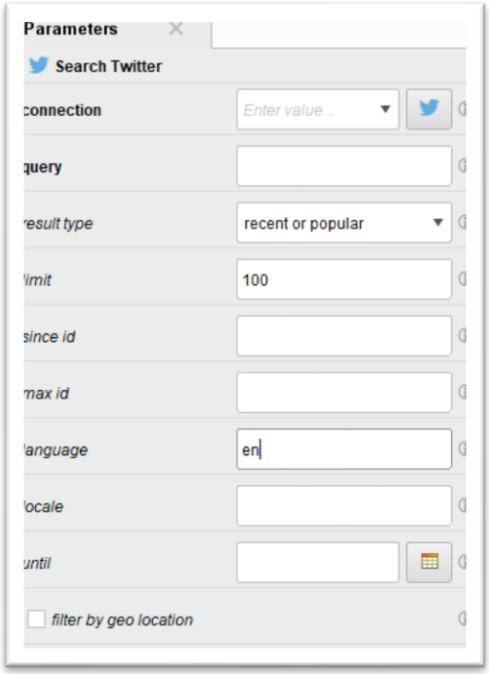
Connection created with Twitter API using Twitter account and Access token.



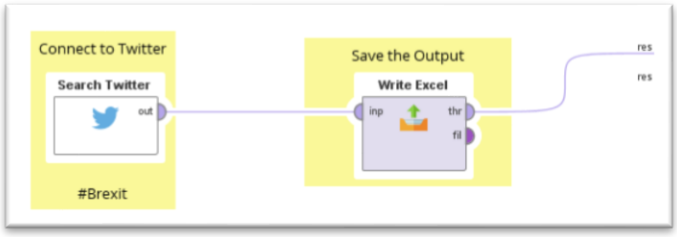

To fetch data from twitter, parameters need to be set to get appropriate data.

The previously created **connection** – Connect Twitter is used to fetch data. Along with that **query** arguments need to be passed which is the keyword twitter API will look for. The keyword searched here is **#Brexit**. The **result type** has 3 options recent, popular, recent or popular. The selected option is recent or popular with the **limit** of 100 to fetch data quicker and most relevant. The **language** chosen is English – en.

Since Twitter API allows last 7 days of data at a time, the process was repeated multiple times to get data. Also, since there were not too many tweets regarding #Brexit, only 1000 samples could be analyzed.



Using **Search Twitter** operator from Aylien plug-in the tweets for the keyword #Brexit were fetched and saved in an excel file.




The data retrieved from Twitter API has the following attributes:

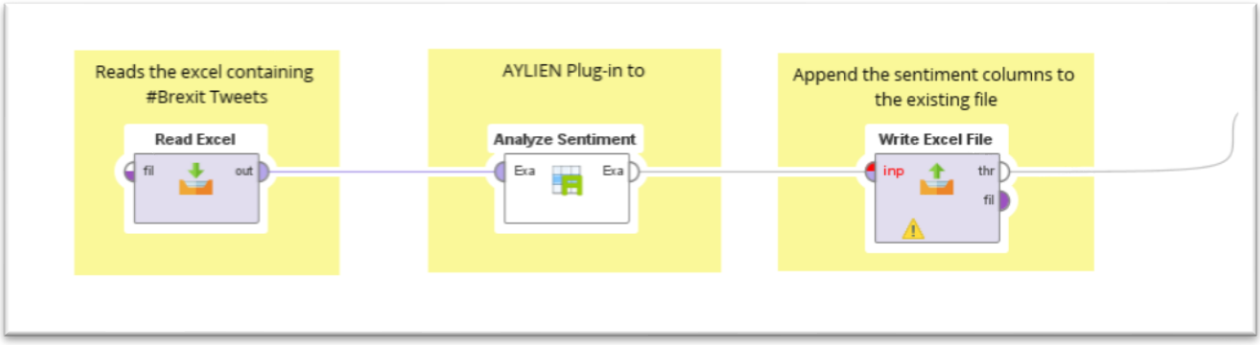
Created At
From User
From User ID
To User
To User ID
Language
Source
Text
Geo Location Latitude
Geo Location Longitude
Retweet Count
ID

Out of all the above attributes the most important one is the **Text** attribute since it contains the tweet which is later used for sentiment analysis.

| Created-At | From-User | From-User-Id | To-User | To-User-Id | Language | Source | Text | Geo-Location... | Geo-Location... | Retweet-Count | Id |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Apr 16, 2019 1... | Nigel Farage | 19017675 | ? | -1 | en | <a href="http://t... | This isn't just a... | ? | ? | 2816 | 111810939778... |
| Apr 15, 2019 1... | Leave Means L... | 776368806553... | ? | -1 | en | <a href="http://t... | John Longwort... | ? | ? | 735 | 111776379103... |
| Apr 15, 2019 3... | Dominic Raab | 4764882552 | ? | -1 | en | <a href="http://t... | UK found to be ... | ? | ? | 1467 | 111779259895... |
| Apr 16, 2019 7... | Zeireen Fuzurally | 33608911 | campbellclaret | 19644592 | en | <a href="http://t... | @campbellclar... | ? | ? | 0 | 111821527741... |
| Apr 16, 2019 7... | Femi Longe | 16718972 | ? | -1 | en | <a href="http://t... | RT @stephen_... | ? | ? | 349 | 111821527560... |
| Apr 16, 2019 7... | Jason | 228300228 | ? | -1 | en | <a href="http://t... | Genuinely men... | ? | ? | 0 | 111821527559... |
| Apr 16, 2019 7... | Euripides Pants | 105663682101... | ? | -1 | en | <a href="https://... | RT @samuel8... | ? | ? | 1 | 111821527513... |
| Apr 16, 2019 7... | D'Arin Moulton | 14481986 | ? | -1 | en | <a href="http://t... | RT @ReutersU... | ? | ? | 3 | 111821527286... |
| Apr 16, 2019 7... | Chris Laughton | 1653832723 | ? | -1 | en | <a href="http://t... | RT @eucopres... | ? | ? | 2685 | 111821527157... |
| Apr 16, 2019 7... | Ambreen | 114031057 | ? | -1 | en | <a href="http://t... | RT @longhurst... | ? | ? | 2 | 111821527051... |
| Apr 16, 2019 7... | james ogram | 851842837 | realDonaldTru... | 25073877 | en | <a href="http://t... | @realDonaldTr... | ? | ? | 0 | 111821527042... |
| Apr 16, 2019 7... | Jim Bishop Gui... | 845896910842... | ? | -1 | en | <a href="http://t... | RT @brexitpart... | ? | ? | 426 | 111821526854... |
| Apr 16, 2019 7... | RGC455 | 107509899410... | ? | -1 | en | <a href="http://t... | RT @spikedonl... | ? | ? | 283 | 111821526828... |
| Apr 16, 2019 7... | Elias M. ??□?... | 306823427 | ? | -1 | en | <a href="http://t... | Farage, Mogg a... | ? | ? | 0 | 111821526599... |
| Apr 16, 2019 7... | Jim Kelly | 105268525 | ? | -1 | en | <a href="http://t... | RT @campbell... | ? | ? | 65 | 111821526506... |
| Apr 16, 2019 7... | John B #FBPE | 2706167375 | ? | -1 | en | <a href="http://t... | RT @snb1969... | ? | ? | 57 | 111821525874... |
| Apr 16, 2019 7... | Simonr1966 | 944437040 | RusbridgeCarl | 2271426747 | en | <a href="https://... | @RusbridgeC... | ? | ? | 0 | 111821525813... |
| Apr 16, 2019 7... | Don | 109416119408... | ? | -1 | en | <a href="http://t... | RT @DrJames... | ? | ? | 50 | 111821525584... |

The excel file is then used to feed data to the **Analyze Sentiment** operator of Aylien plug-in. The output of the operator is saved in the same excel file. The operator adds the following 4 columns to the existing file,

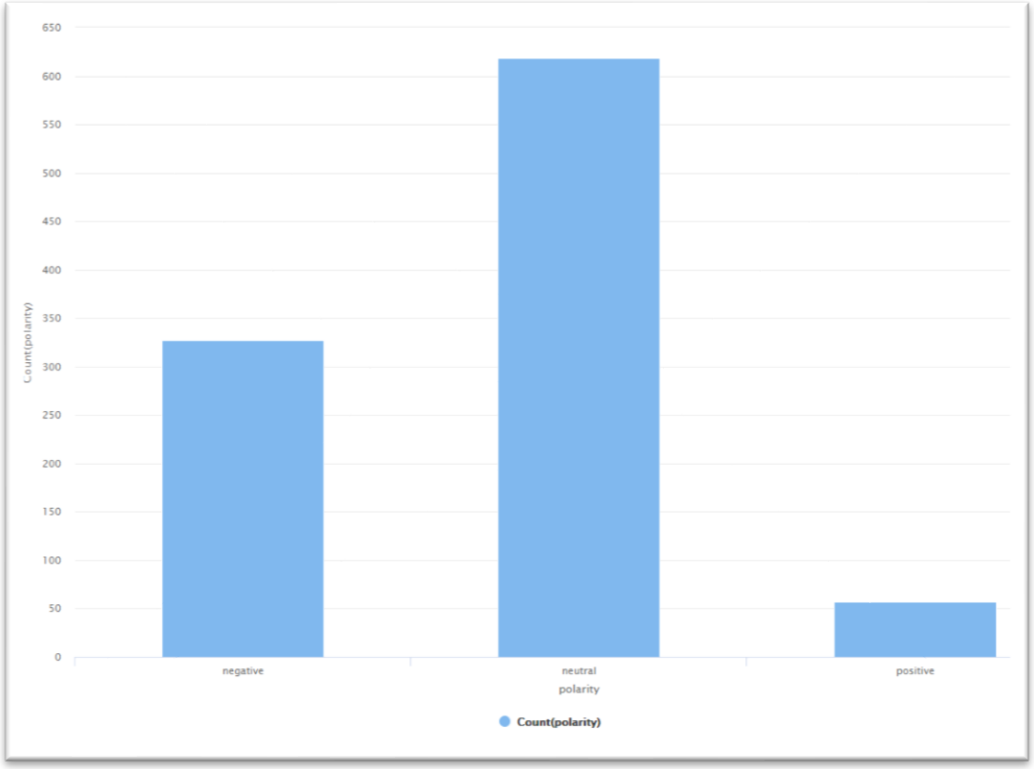
- polarity – positive, negative or neutral
- subjectivity – subjective (personal opinion) or objective (fact)
- polarity_confidence – confidence score of polarity
- subjectivity_confidence - confidence score of subjectivity




| Text<br>Category | polarity<br>Category | subjectivity<br>Category | polarity_confidence<br>Number | subjectivity_confidence<br>Number |
|---|---|---|---|---|
| This isn't just about Brexit anymore, it's about what ki... | negative | subjective | 0.958 | 1 |
| John Longworth announces he will stand for The Brex... | neutral | subjective | 0.789 | 1 |
| UK found to be hottest investment destination despite... | neutral | objective | 0.779 | 1.000 |
| @campbellclaret @Jeremy_Hunt Actually wondered t... | neutral | subjective | 0.558 | 1 |
| RT @stephen_dorrell: Have finally reactivated my Twit... | neutral | subjective | 0.942 | 0.904 |
| Genuinely mental that MP's who are running the count... | positive | objective | 0.359 | 1.000 |
| RT @samuel88davies: @JimMFelton Wasn't it Will Se... | negative | subjective | 0.696 | 1 |
| RT @ReutersUK: Sterling slips on Brexit talks concer... | neutral | objective | 0.981 | 1.000 |
| RT @eucopresident: At the summit one Prime Ministe... | neutral | subjective | 0.648 | 1 |
| RT @longhurstkerry: @Nigel_Farage Can you imagin... | neutral | subjective | 0.795 | 1 |
| @realDonaldTrump dear president trump will you co... | neutral | subjective | 0.584 | 1 |
| RT @brexitparty_uk: Join @Nigel_Farage @TiceRich... | neutral | objective | 0.867 | 1.000 |
| RT @spikedonline: Brexit is the opposite of Nazism. It... | negative | subjective | 0.478 | 1 |
| Farage, Mogg and Johnson: The Phoney Defenders o... | neutral | objective | 0.823 | 1 |
| RT @campbellclaret: You inspired this piece @jerem... | negative | subjective | 0.955 | 1 |
| RT @snb19692: Is this what we are fighting for? | neutral | subjective | 0.978 | 1 |
| @RusbridgeCarl But they have got a 100,000 membe... | negative | subjective | 0.605 | 1 |
| RT @DrJamesKent3: I am aware of how my followers... | negative | subjective | 0.959 | 1 |
| RT @MillyMia8: @Doozy_45 BS - Milne wants brexit th... | neutral | objective | 0.613 | 1 |

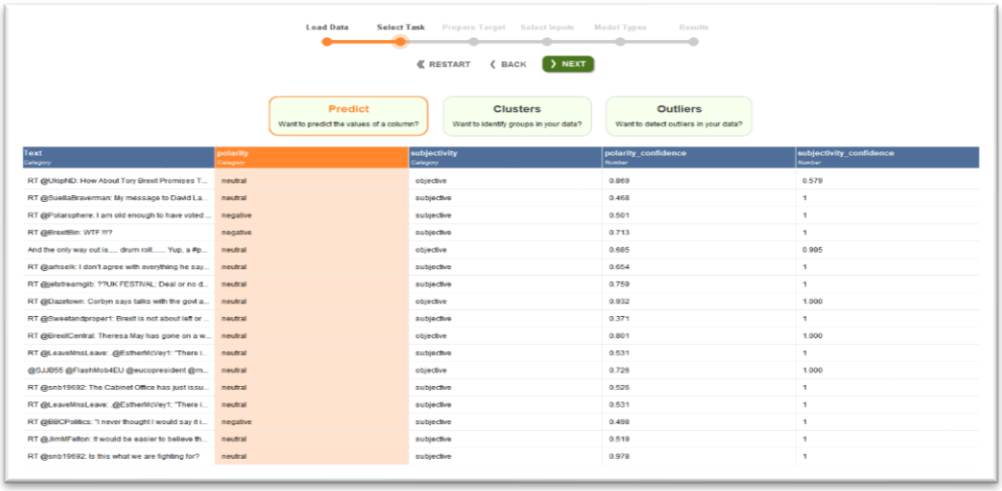
## Data Preparation

Now the sentiment of tweets are identified. The following figure shows that the majority of tweets are neutral in nature followed by negative tweets and very low contribution of positive tweets.




The Aylien plug-in helps us to identify the sentiment but it does not explain which Machine Learning Algorithm(MLA) is used for the purpose. So to identify the MLA auto-model is run. It will help us to know which MLA can perform the best for such a task. Now, since we have the label for the tweets we can run the auto-model to **Predict** the label(polarity) for the tweet.

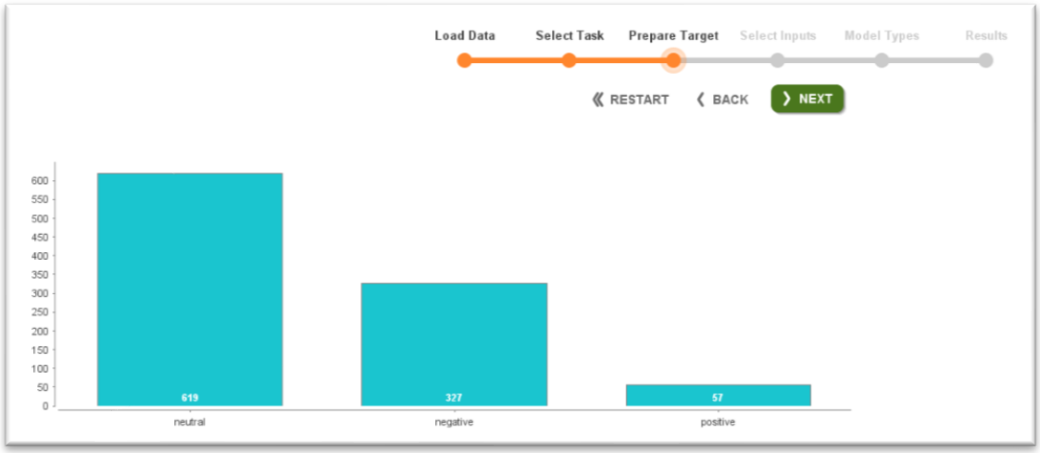
The data preparation for this task is to remove all the attributes which add no value to the sentiment analysis. All the attributes retrieved from Twitter API **except Text** are removed. Now only 5 attributes are considered as shown in the image below.




| Text | polarity | subjectivity | polarity_confidence | subjectivity_confidence |
|---|---|---|---|---|
| RT @UkipND: How About Tory Brexit Promises T... | neutral | objective | 0.869 | 0.579 |
| RT @SuellaBraverman: My message to David La... | neutral | subjective | 0.468 | 1 |
| RT @Polarsphere: I am old enough to have voted ... | negative | subjective | 0.501 | 1 |
| RT @BrextBin: WTF !!? | negative | subjective | 0.713 | 1 |
| And the only way out is..... drum roll..... Yup, a #p... | neutral | objective | 0.685 | 0.985 |
| RT @amselk: I don't agree with everything he say... | neutral | subjective | 0.654 | 1 |
| RT @jetstreamgib: ??UK FESTIVAL: Deal or no d... | neutral | subjective | 0.759 | 1 |
| RT @Dazetown: Corbyn says talks with the govt a... | neutral | objective | 0.932 | 1.000 |
| RT @Sweetandproper1: Brexit is not about left or ... | neutral | subjective | 0.371 | 1 |
| RT @BrexitCentral: Theresa May has gone on a w... | neutral | objective | 0.801 | 1.000 |
| RT @LeaveMnsLeave: @EstherMcVey1: "There i... | neutral | subjective | 0.531 | 1 |
| @GJJ855 @FlashMob4EU @eucopresident @m... | neutral | objective | 0.726 | 1.000 |
| RT @snb19692: The Cabinet Office has just issu... | neutral | subjective | 0.526 | 1 |
| RT @LeaveMnsLeave: @EstherMcVey1: "There i... | neutral | subjective | 0.531 | 1 |
| RT @BBCPolitics: "I never thought I would say it i... | negative | subjective | 0.498 | 1 |
| RT @JimMFalton: It would be easier to believe th... | neutral | subjective | 0.519 | 1 |
| RT @snb19692: Is this what we are fighting for? | neutral | subjective | 0.978 | 1 |

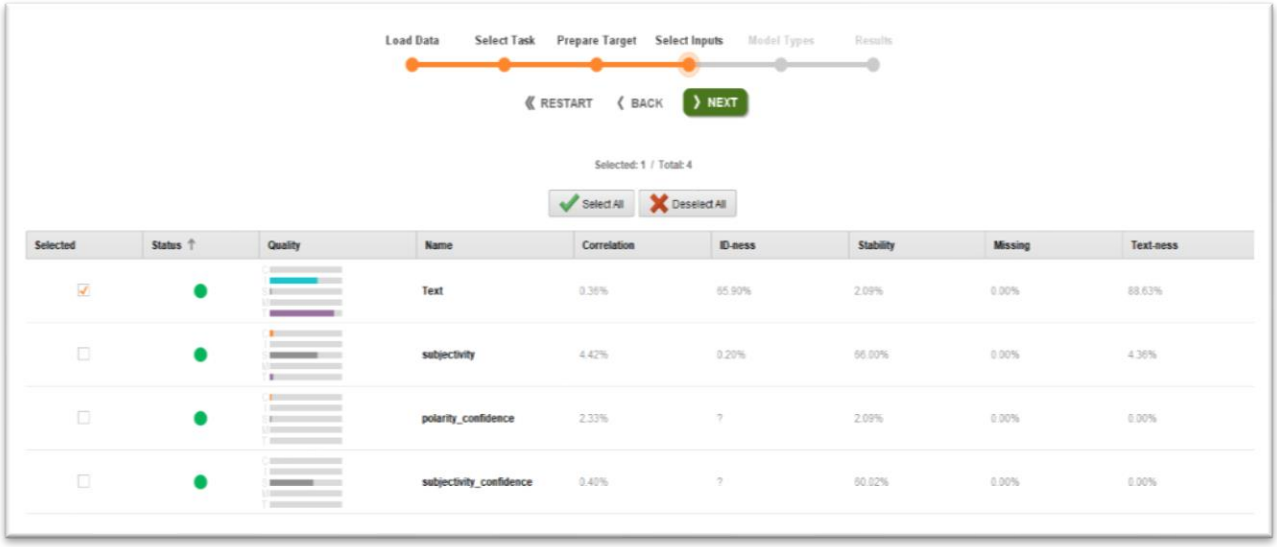
The MLA needs to classify the Text(tweet) into categories given by polarity. Hence, polarity is the target variable/label.

The below image shows the contribution of each category in the label.

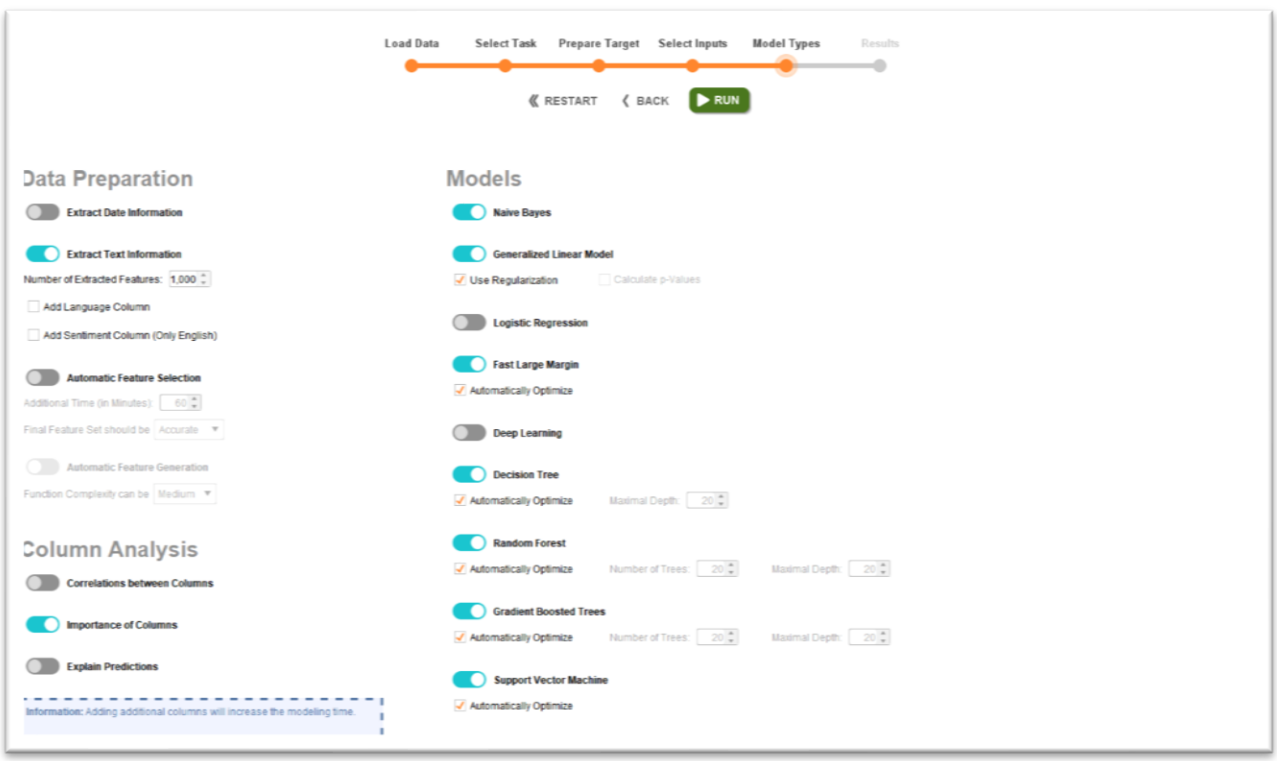


# Modelling

To train the model appropriate input needs to be fed. The auto-model suggests that all the attributes are good and can be used to predict the label. But since we need the sentiment based only on text, we deselect all other attributes.



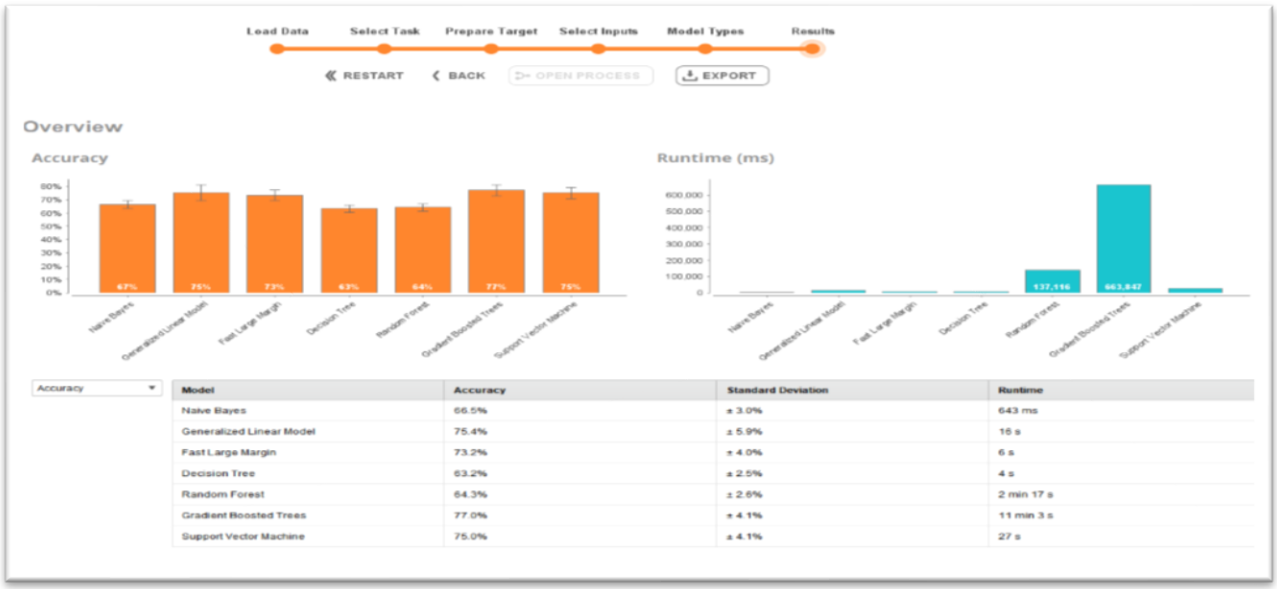

The auto-model suggests all the classification models which can perform best on the given dataset. Without changing anything we go ahead with the suggested settings.



# Evaluation

The model runs and gives us the comparison and detailed results of all the algorithms.

It is evident from the comparison below that the algorithm with least runtime, acceptable accuracy and comparatively stable is Naïve Bayes.




The performance table of Naïve Bayes shows how accurate it was to identify each category. Overall accuracy being 66.54% which is not very bad.


| | true negative | true neutral | true positive | class precision |
|---|---|---|---|---|
| pred. negative | 67 | 42 | 2 | 60.36% |
| pred. neutral | 30 | 116 | 6 | 76.32% |
| pred. positive | 2 | 14 | 8 | 33.33% |
| class recall | 67.68% | 67.44% | 50.00% | |

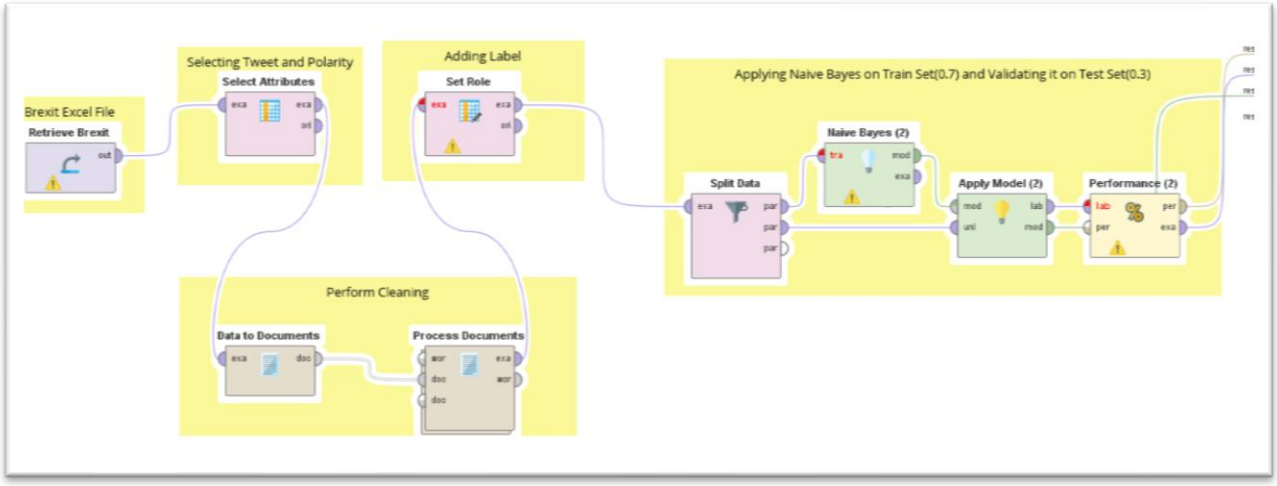
# Deployment

## Using Rapid Miner Studio – Manual Model

Now that we know the most suitable MLA for the given dataset and task at hand.

Let's design a manual model to train and test the model. Use the excel file which contains the required columns. **Select Attributes** operator helps to select only the important attributes in this case, text and polarity.

To perform cleaning, operator **Data to documents** causes the next operator require document as input and **Process documents** are used.




Process documents contains various operators which needs to be applied on the document data.
**Transform cases** – to convert all the text to lower case

**Tokenize** – to split the text of a document into a sequence of tokens

**Replace tokens** – to remove punctuation marks, numbers from the text

**Filter StopWords** – to filter English stopwords from a document by removing every token which equals a stopword from the built-in stopword list.

**Generate n-grams** – to create term n-Grams of tokens in a document. It is defined as a series of consecutive tokens of length n.


The output of Process document operator is fed to **Set Role** operator to define attribute Polarity as label for further process.

Data is then split into trainset(0.7) and testset(0.3). The trainset is fed to the **Naïve Bayes** operator for training the model. The testset is fed to the **Apply model** operator along with the learnings of Naïve Bayes operator.

To know the performance of the model **Performance** operator is used.

The following figures show the accuracy of the manual model. The accuracy achieved (78.74%) is much better than the auto-model (66.54%).

```
PerformanceVector:
accuracy: 78.74%
ConfusionMatrix:
True:   negative        neutral positive
negative:       45      0       0
neutral:        53      186     11
positive:       0       0       6
```

The class precision achieved by manual model is way better than auto-model.

**accuracy: 78.74%**

| | true negative | true neutral | true positive | class precision |
|---|---|---|---|---|
| pred. negative | 45 | 0 | 0 | 100.00% |
| pred. neutral | 53 | 186 | 11 | 74.40% |
| pred. positive | 0 | 0 | 6 | 100.00% |
| class recall | 45.92% | 100.00% | 35.29% | |

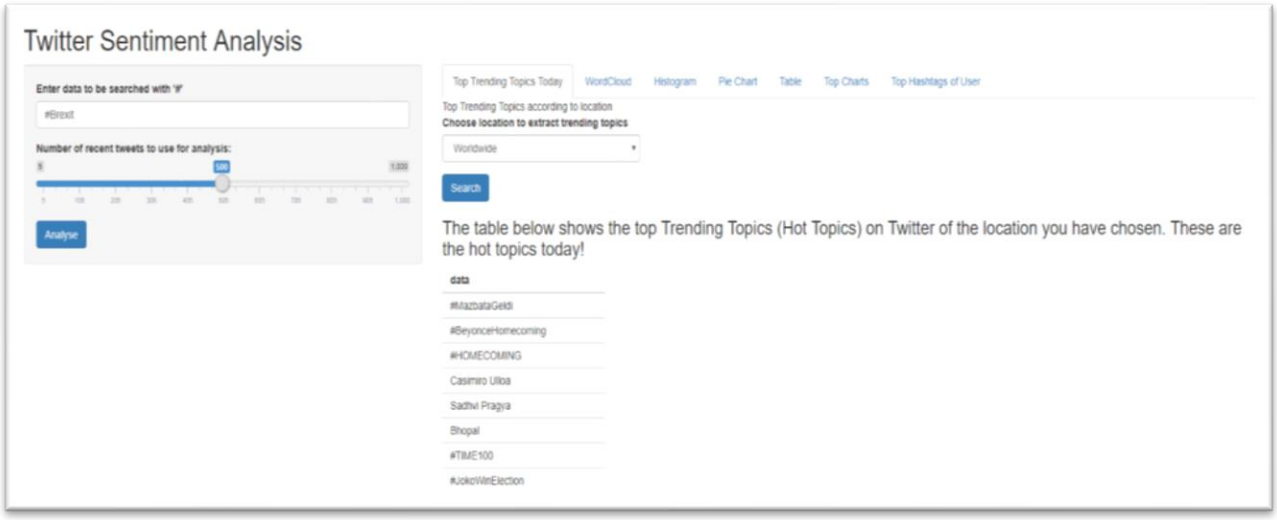
## Using R and Shiny

A WebApp is developed to deploy the Twitter Sentiment Analysis project. The aim for the deployment is to connect non-technical people to the field of data mining. The simple user interface will help user in creating a report based on keyword search. The application is very user friendly.

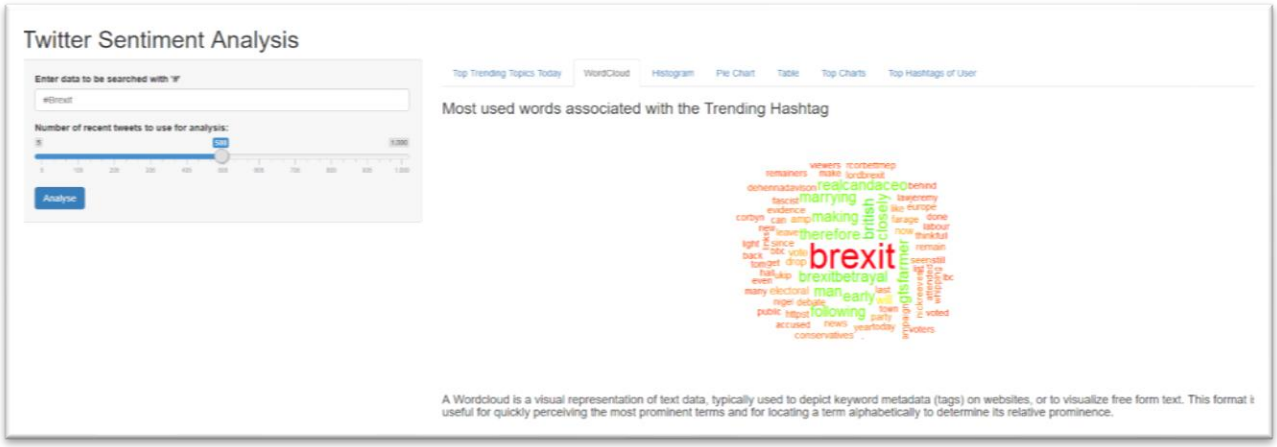
The WebAapp has a search textbox to search for a specific keyword. Enter the keyword along with hashtag to find tweets related to it. Users can mention number of tweets to be analyzed between 5 and 1000.

By default the search is worldwide but location can be changed using the drop-down menu.

**Top Trending Topics Today** – Trending topics are listed for the selected location. The list indicates the hot topics which user can search for, to get good amount of tweets and better analysis.




**Word Cloud**: It is used as a visual representation of text data. The format is useful for faster and to easily perceive prominent textual data. The figure below shows the word cloud for #Brexit.



**Histogram** – It graphically depicts the opinions of people about #Brexit.


**Pie Chart –** A pie chart is a circular statistical graphic, which is divided into slices to illustrate the sentiment of the hashtag.

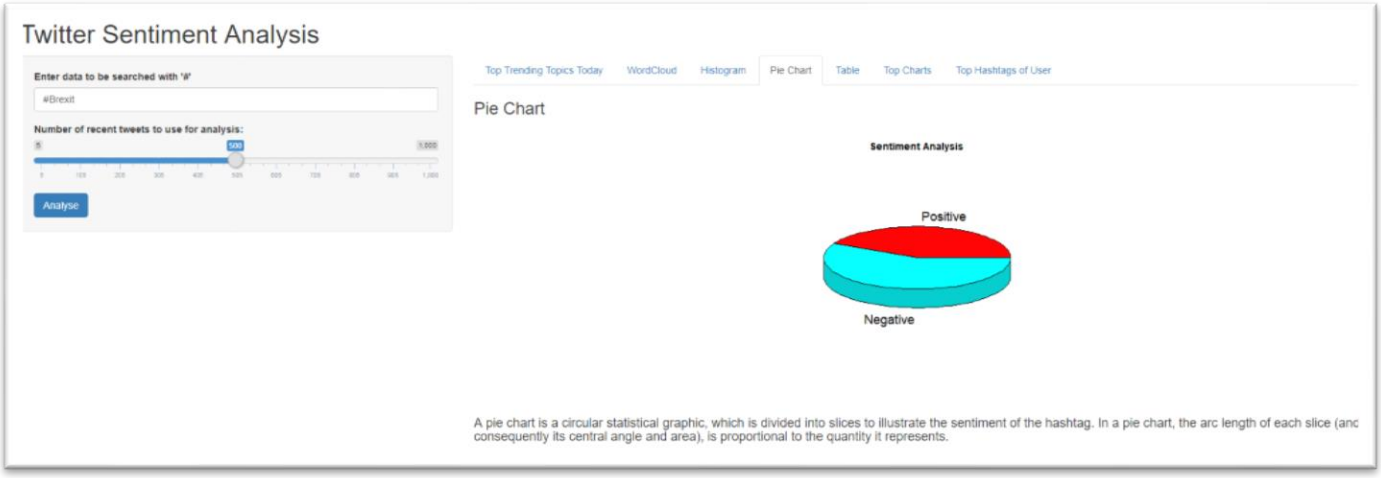
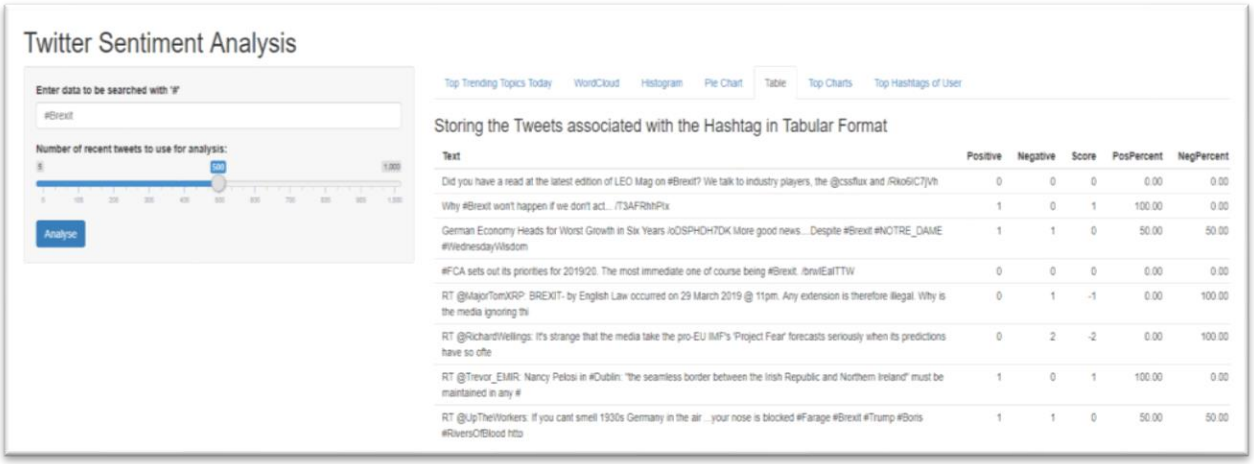
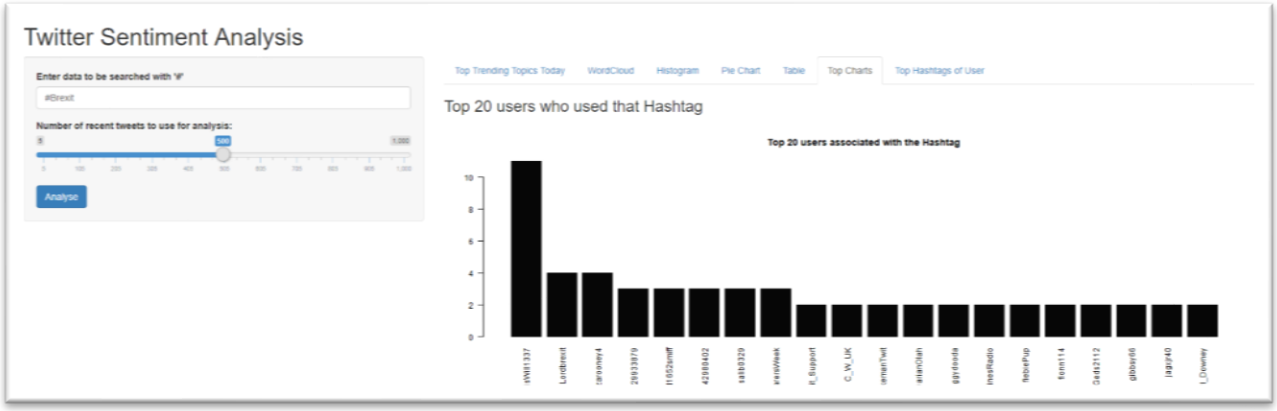


**Table –** It depicts the Sentiment of the Tweets (Positive, Negative or Neutral) associated with the searched Hashtag by showing the score for each type of sentiment.

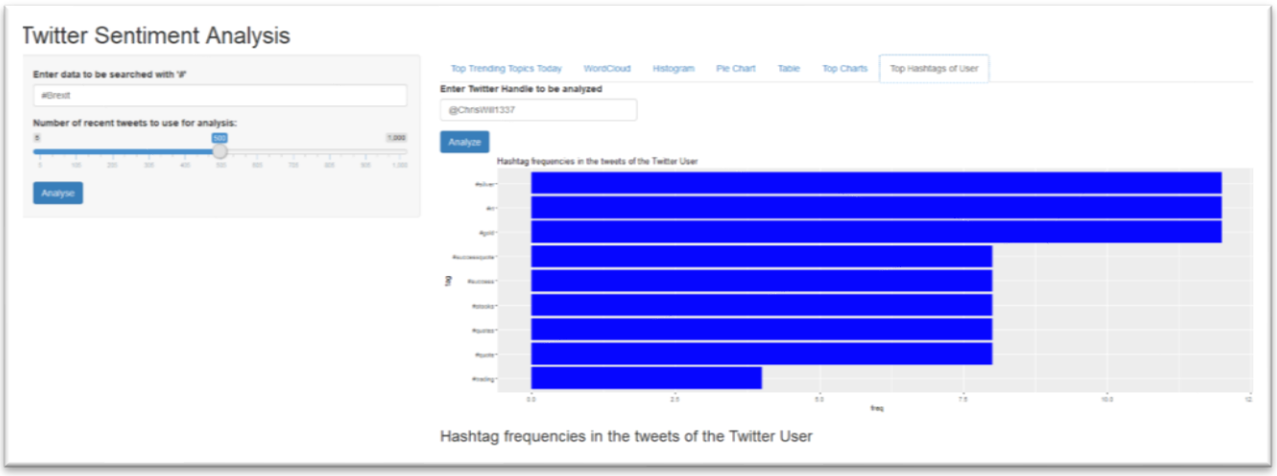



| Text | Positive | Negative | Score | PosPercent | NegPercent |
|---|---|---|---|---|---|
| Did you have a read at the latest edition of LEO Mag on #Brexit? We talk to industry players, the @cssflux and /Rko6iC7jVh | 0 | 0 | 0 | 0.00 | 0.00 |
| Why #Brexit won't happen if we don't act... /T3AFRhtPix | 1 | 0 | 1 | 100.00 | 0.00 |
| German Economy Heads for Worst Growth in Six Years /oOSPHDH7DK More good news...Despite #Brexit #NOTRE_DAME #WednesdayWisdom | 1 | 1 | 0 | 50.00 | 50.00 |
| #FCA sets out its priorities for 2019/20. The most immediate one of course being #Brexit. /brwIEalTTW | 0 | 0 | 0 | 0.00 | 0.00 |
| RT @MajorTomXRP: BREXIT- by English Law occurred on 29 March 2019 @ 11pm. Any extension is therefore illegal. Why is the media ignoring thi | 0 | 1 | -1 | 0.00 | 100.00 |
| RT @RichardWellings: It's strange that the media take the pro-EU IMF's 'Project Fear' forecasts seriously when its predictions have so ofte | 0 | 2 | -2 | 0.00 | 100.00 |
| RT @Trevor_EMIR: Nancy Pelosi in #Dublin: "the seamless border between the Irish Republic and Northern Ireland" must be maintained in any # | 1 | 0 | 1 | 100.00 | 0.00 |
| RT @UpTheWorkers: If you cant smell 1930s Germany in the air ... your nose is blocked #Farage #Brexit #Trump #Boris #RiversOfBlood http | 1 | 1 | 0 | 50.00 | 50.00 |

**Top Charts** – It shows the top 20 users who used the searched text - #Brexit




**Top Hashtag of User** – In this tab, user can check the top hashtags used by any given twitter handle. Here we picked the twitter handle @ChrisWill1337 who used #Brexit the most.



# Conclusion

This Data Mining assignment has helped us to learn and explore Rapid Miner, R Shiny and Text Analysis concepts in an interesting and fun way. The analysis for #Brexit gives shocking results of having neutral sentiment followed by positive and very little negative. We learn that the process of sentiment analysis gives us deeper insights regarding a specific topic.

# References

https://www.bbc.com/news/uk-46318565

https://developer.twitter.com/content/developer-twitter/en.html

http://blog.aylien.com/building-a-twitter-sentiment-analysis-process-in/

https://shiny.rstudio.com/tutorial/