

**DUBLIN BUSINESS SCHOOL**

# **APPLICATION OF DATA MINING TOOLS&TECHNIQUES**

**DATA MINING - B9DA103**

**TERRI HOARE**

Nishad Abdul latheef  
Student ID: 10382242  
Date: 8-14-2018

## Contents

Introduction .....	3
Business Understanding.....	4
Data Understanding.....	5
Data Preparation.....	6
Modelling .....	7
Evaluation .....	10
Deployment .....	11
Conclusion.....	15
References .....	17

# **Introduction**

Data mining is defined as the process of exploring through huge data sets to discover patterns and create relationships that helps in solving problems by the way of data analysis. The modern data mining tools enables various enterprises to predict future trends.

In data mining, Text mining is defined as the process of drilling through and analysing huge amounts of unstructured text data with the help of software which can discover topics, concepts, keywords, patterns, and other relevant attributes in the data.

Twitter is one of the most popular social networking sites used for microblogging. Each tweets are restricted to 280 characters. All over the world, twitter is used as a platform to convey one's emotions. Some companies even target on analysing these emotions on a specific topic for various purposes. The aim of the assignment is to build a model to detect and summarise the sentiments using the tweets on random topics.

The whole process of this analysis was carried out with the CRISP-DM data modelling steps in mind. CRISP-DM is implemented by splitting the whole data mining process into six steps. These steps are given as;

- business understanding,
- data understanding,
- data preparation,
- modelling,
- evaluation, and
- deployment.

These steps aid an organization to figure out the data mining process and provide a schema to follow while planning and executing a data mining work.

# **Business Understanding**

The model that we created is a tool to analyse sentiment from the given text. The sentiment analysis is a text mining technique to extract, identify and study the polarity of a given text or document. The results show whether the emotion conveyed in the document is positive, or negative. In principle, the accuracy of a sentiment analysis system is determined by how well it finds similarity with human judgments. This is usually calculated by different measures based on precision and recall over the two target variables of negative and positive texts.

From a business point of view, the output of the data could be utilised for various purposes. This could be for political campaigns, social crisis management or purely for business related goals. Some of the benefits of sentiment analysis in real world are:

- Sentiment Analysis in BI setup is a key application. Based on the reviews generated through sentiment analysis in business, you can always adjust to the present market situation and satisfy your customers in a better way.
- Knowing the sentiment data of your competitors gives you the opportunity as well as the incentive to perk up your performance.
- The tone and temperament of customers experience data can be detected and then categorized according to the sentiments attached. This helps to know what is being properly implemented with regards to products, services and customer support and what needs improvement.
- The polarity of the tweeters towards a particular government can be used to predict the election result

# Data Understanding

The dataset was obtained from the site '<http://help.sentiment140.com/for-students/>'. It includes a link to a CSV file (Fig 1) which contains twitter data of random users on random topics.

File	Home	Insert	Draw	Page Layout	Formulas	Data	Review	View	Help	Tell me what you want to do
A1										
	A	B	C	D	E	F				
1	Sentiment	ID	Date	Query	Username	Tweets				
2	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by texting it... and might cry as a result School tr				
3	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Managed to save 50% The rest go out of bounc				
4	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire				
5	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all. i'm mad. why am i here? because I can't see \				
6	0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kweseidei not the whole crew				
7	0	1467811592	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	mybirch	Need a hug				
8	0	1467811594	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	coZZ	@LOLTrish hey long time no see! Yes... Rains a bit ,only a bit LOL , I'm fine thanks , how's y				
9	0	1467811795	Mon Apr 06 22:20:05 PDT 2009	NO_QUERY	2Hood4Hollywood	@Tatiana_K nope they didn't have it				
10	0	1467812025	Mon Apr 06 22:20:09 PDT 2009	NO_QUERY	mimismo	@twittera que me muera ?				
11	0	1467812416	Mon Apr 06 22:20:16 PDT 2009	NO_QUERY	erinx3leanexo	spring break in plain city... it's snowing				
12	0	1467812579	Mon Apr 06 22:20:17 PDT 2009	NO_QUERY	pardonlauren	I just re-pierced my ears				
13	0	1467812723	Mon Apr 06 22:20:19 PDT 2009	NO_QUERY	TLeC	@caregiving I couldn't bear to watch it. And I thought the UA loss was embarrassing . . . .				
14	0	1467812771	Mon Apr 06 22:20:19 PDT 2009	NO_QUERY	robobbierobert	@octolin216 It it counts, idk why I did either. you never talk to me anymore				
15	0	1467812784	Mon Apr 06 22:20:20 PDT 2009	NO_QUERY	bayofwolves	@smarrison i would've been the first, but i didn't have a gun. not really though, zac snyde				
16	0	1467812799	Mon Apr 06 22:20:20 PDT 2009	NO_QUERY	HairByJless	@iamjazzyfizzle I wish I got to watch it with you!! I miss you and @iamlinicki how was the				
17	0	1467812964	Mon Apr 06 22:20:22 PDT 2009	NO_QUERY	lovesongwriter	Hollis' death scene will hurt me severely to watch on film wry is directors cut not out now?				
18	0	1467813137	Mon Apr 06 22:20:25 PDT 2009	NO_QUERY	armotley	about to file taxes				
19	0	1467813579	Mon Apr 06 22:20:31 PDT 2009	NO_QUERY	starkissed	@LettyA ahh ive always wanted to see rent love the soundtrack!!				
20	0	1467813782	Mon Apr 06 22:20:34 PDT 2009	NO_QUERY	gi_gi_bee	@FakerPattyPattz Oh dear. Were you drinking out of the forgotten table drinks?				
21	0	1467813985	Mon Apr 06 22:20:37 PDT 2009	NO_QUERY	quanvu	@alydesigns i was out most of the day so didn't get much done				
22	0	1467813992	Mon Apr 06 22:20:38 PDT 2009	NO_QUERY	swinspeedx	one of my friend called me, and asked to meet with her at Mid Valley today...but i've no tin				
23	0	1467814119	Mon Apr 06 22:20:40 PDT 2009	NO_QUERY	cooliodoc	@angry_barista I baked you a cake but I ated it				
24	0	1467814180	Mon Apr 06 22:20:40 PDT 2009	NO_QUERY	villLLante	this week is not going as i had hoped				
25	0	1467814192	Mon Apr 06 22:20:41 PDT 2009	NO_QUERY	Ljelli3166	blagh class at 8 tomorrow				
26	0	1467814438	Mon Apr 06 22:20:44 PDT 2009	NO_QUERY	ChicagoCubbie	I hate when I have to call and wake people up				
27	0	1467814783	Mon Apr 06 22:20:47 PDT 2009	NO_QUERY	thebigbuck	@thebigbuck I'm not a fan of the movie but I love the soundtrack!!				

Fig 1

The training set is an excel file that contains 200,000 rows and 6 columns. The six fields of the training set are:

1. the polarity of the tweet (0 = negative and 4 = positive).
2. the id of the tsweet.
3. the date of the tweet.
4. the query. If there is no query, then this value is NO\_QUERY.
5. the user that tweeted.
6. the text of the tweet.

A test set with 50,000 rows of tweets and emotions are used for predicting the sentiment

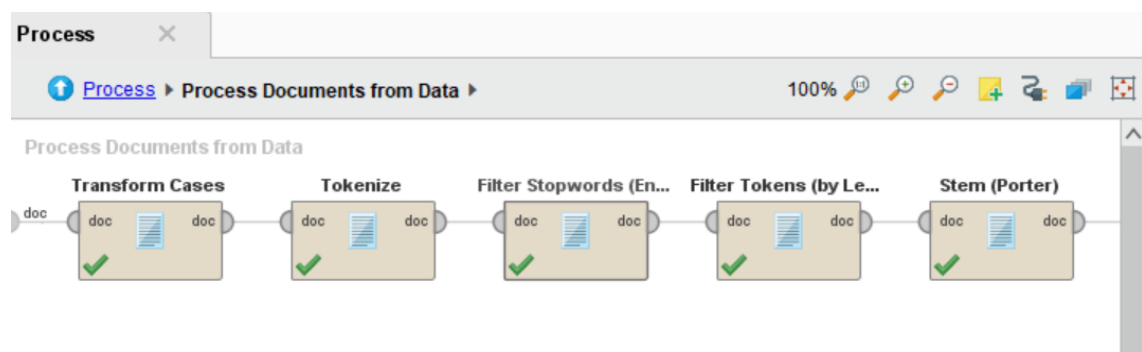
# Data Preparation

The dataset obtained from the site contains tweet text that needs filtering and preparing for the main process. This is done by the operators in RapidMiner named as:

- **Nominal to Text Operator**

The type of selected nominal attributes is changed to text before it is processed in process document operator.

- **Process Document Operator**



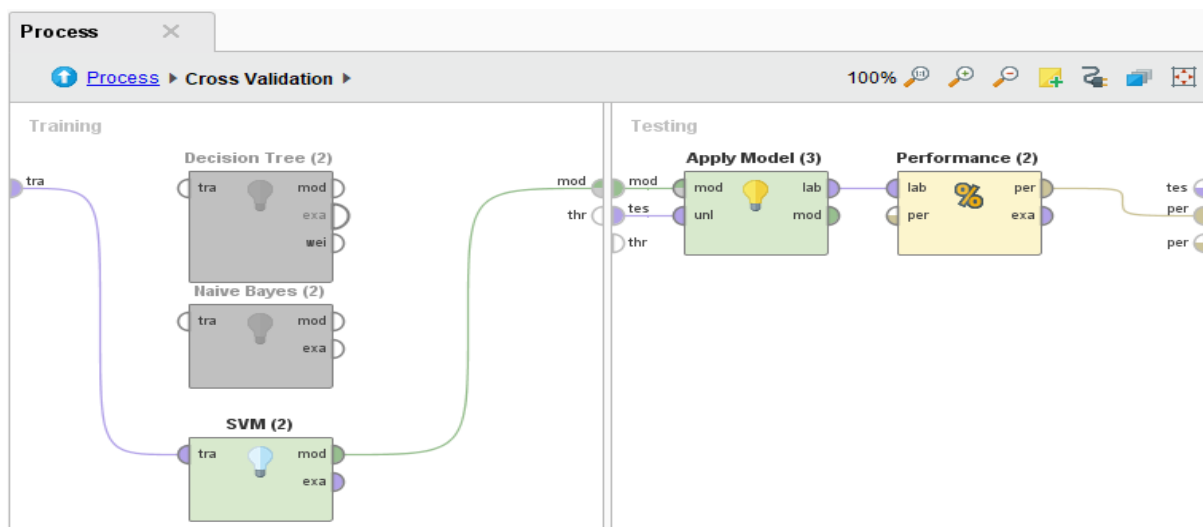
**Fig 2**

The process document operator (Fig 2) consists of the following steps:

1. **Transform Cases** - Transforms cases of characters in a document
2. **Tokenize** - splits the text of a document into a sequence of tokens
3. **Filter Stopwords** - filters English stop words from a document by removing every token matching built-in stop-word list.
4. **Filter Tokens** - Filters tokens based on their length
5. **Stem (Porter)** - stems English words using the Porter stemming algorithm that reduce the length of the words until a minimum length is reached

# Modelling

In the modelling phase, various modelling techniques were applied to our dataset. The performance and quality of each model were assessed after exploring with different models on the same dataset. Finally, the best and the most apt technique was chosen to create a final model for the chosen data mining problem in hand.



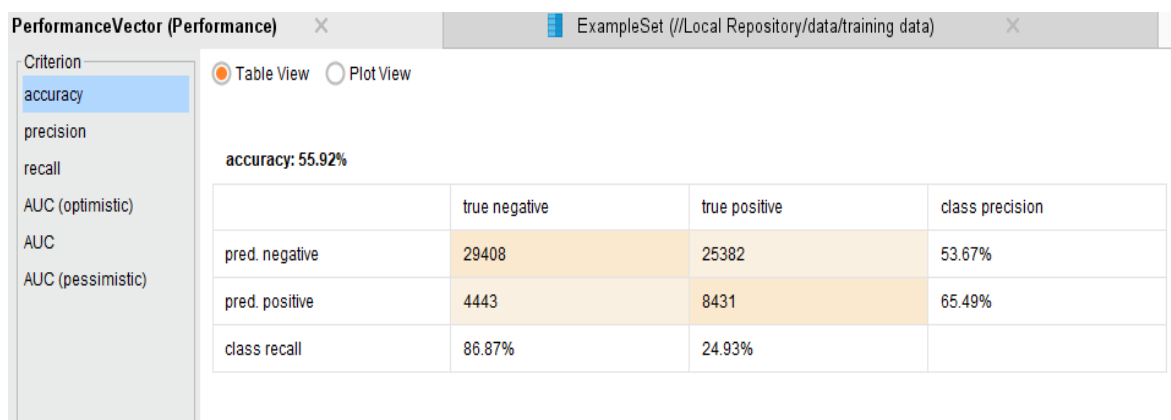
**Fig 3**

Three major varieties of machine learning algorithms were used for this sentiment analysis to train the model in which the label was set. The cross-validation operator in RapidMiner (Fig 3) was used to design these algorithms. Since the cross-validation operator is more effective in training a text model, the particular validation is used here. The first step of validation is partitioning the given dataset into k subsets of equal size. One of the subsets is then chosen as the test data set and the remaining k - 1 subsets are utilised as training data. The validation process is then repeated k times, with each of the k subsets used exactly once as the test data. The average of the k results from the k iterations are used to produce a single estimation.

The cross-validation operator consists of three various algorithms namely SVM, Naïve Bayes and Decision Tree. All three algorithms are applied to train the model and the calculated accuracy is checked by comparing the results from applying the model to the test data. The accuracy of each algorithm is produced by the performance operator so that the user can decide on which methodology to use for further steps

The performance of various modelling techniques to do the sentiment analysis for our dataset are:

- Decision tree



PerformanceVector (Performance) ExampleSet (//Local Repository/data/training data)

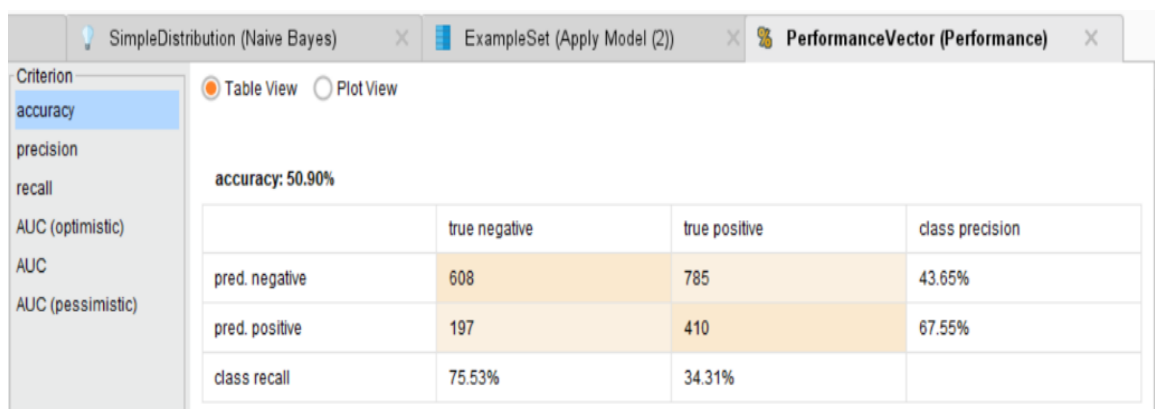
Table View Plot View

accuracy: 55.92%

	true negative	true positive	class precision
pred. negative	29408	25382	53.67%
pred. positive	4443	8431	65.49%
class recall	86.87%	24.93%	

Fig 4

- Naive Bayes



SimpleDistribution (Naive Bayes) ExampleSet (Apply Model (2)) PerformanceVector (Performance)

Table View Plot View

accuracy: 50.90%

	true negative	true positive	class precision
pred. negative	608	785	43.65%
pred. positive	197	410	67.55%
class recall	75.53%	34.31%	

Fig 5



- SVM

The screenshot shows the 'PerformanceVector (Performance)' window in Weka. The 'Criterion' list on the left includes accuracy, precision, recall, AUC (optimistic), AUC, and AUC (pessimistic). The 'Table View' radio button is selected. The overall accuracy is 66.10%.

	true negative	true positive	class precision
pred. negative	234	107	68.62%
pred. positive	571	1088	65.58%
class recall	29.07%	91.05%	

**Fig 6**

# Evaluation

This step of Evaluation involves determining if some crucial business issue has not been adequately considered. At the end of this step, we had to determine exactly how to make use of this data mining results.

From the results of performance operators, SVM (Fig 6) had the best accuracy rate. So, this model is selected for further steps. support vector machines are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis.

With an accuracy of 66.10%, the model has correctly labelled the tweets from the test dataset as positive, negative or neutral (Fig 7). This is validated on a separate test dataset of 5000 tweets to verify the accuracy of prediction.

ExampleSet (2450 examples, 4 special attributes, 17215 regular attributes) Filter (2,450 / 2,450 examples): all

Row No.	Sentiment	prediction(S...	confidence(...	confidence(...	aa	aaa	aaaaaa	aaaand	aaaa
1	negative	negative	0.563	0.437	0	0	0	0	0
2	negative	negative	0.514	0.486	0	0	0	0	0
3	negative	positive	0.461	0.539	0	0	0	0	0
4	negative	positive	0.456	0.544	0	0	0	0	0
5	negative	positive	0.497	0.503	0	0	0	0	0
6	negative	positive	0.474	0.526	0	0	0	0	0
7	negative	negative	0.628	0.372	0	0	0	0	0
8	positive	positive	0.408	0.592	0	0	0	0	0
9	positive	positive	0.484	0.516	0	0	0	0	0
10	negative	positive	0.340	0.660	0	0	0	0	0

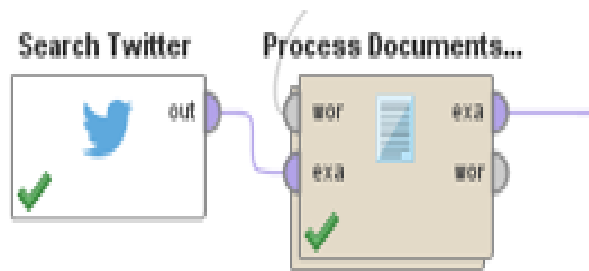
Fig 7

The accuracy level suggests that the model can be used effectively as a solution to the suggested business problem and this will be applied in the deployment phase.

# **Deployment**

The final phase deals with the implementation of the solution across the problem. The implementation can be of different forms like generating a report, a dashboard or a visualization, an application or website or even complete integration to the business.

In this case, the deployment phase we selected involves the applying of the obtained solution on datasets of tweets on leaders of different countries and comparing the results. In order to do that, a total of 2000 tweets on different leaders were collected using the search twitter operator in RapidMiner

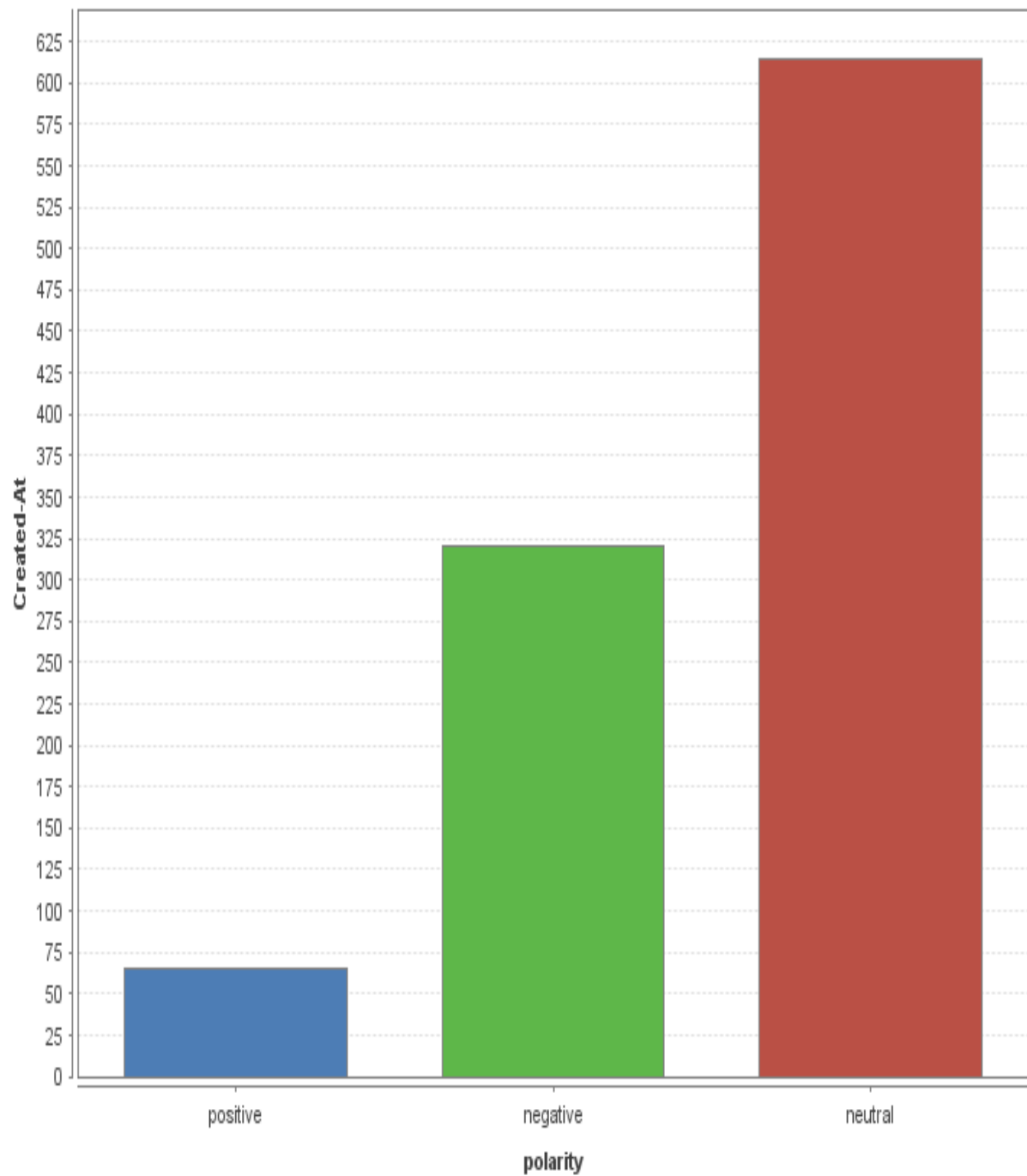


**Fig 8**

The first dataset contains 1000 tweets on Mr. Donald Trump, the current President of the United States, the second dataset is 500 tweets on Mr. Barack Obama, former President of the United States and the final dataset comprises of 500 tweets on Mr. Narendra Modi, the current Prime Minister of India.

Following are the results obtained from applying the sentiment analysis model on these 3 datasets.

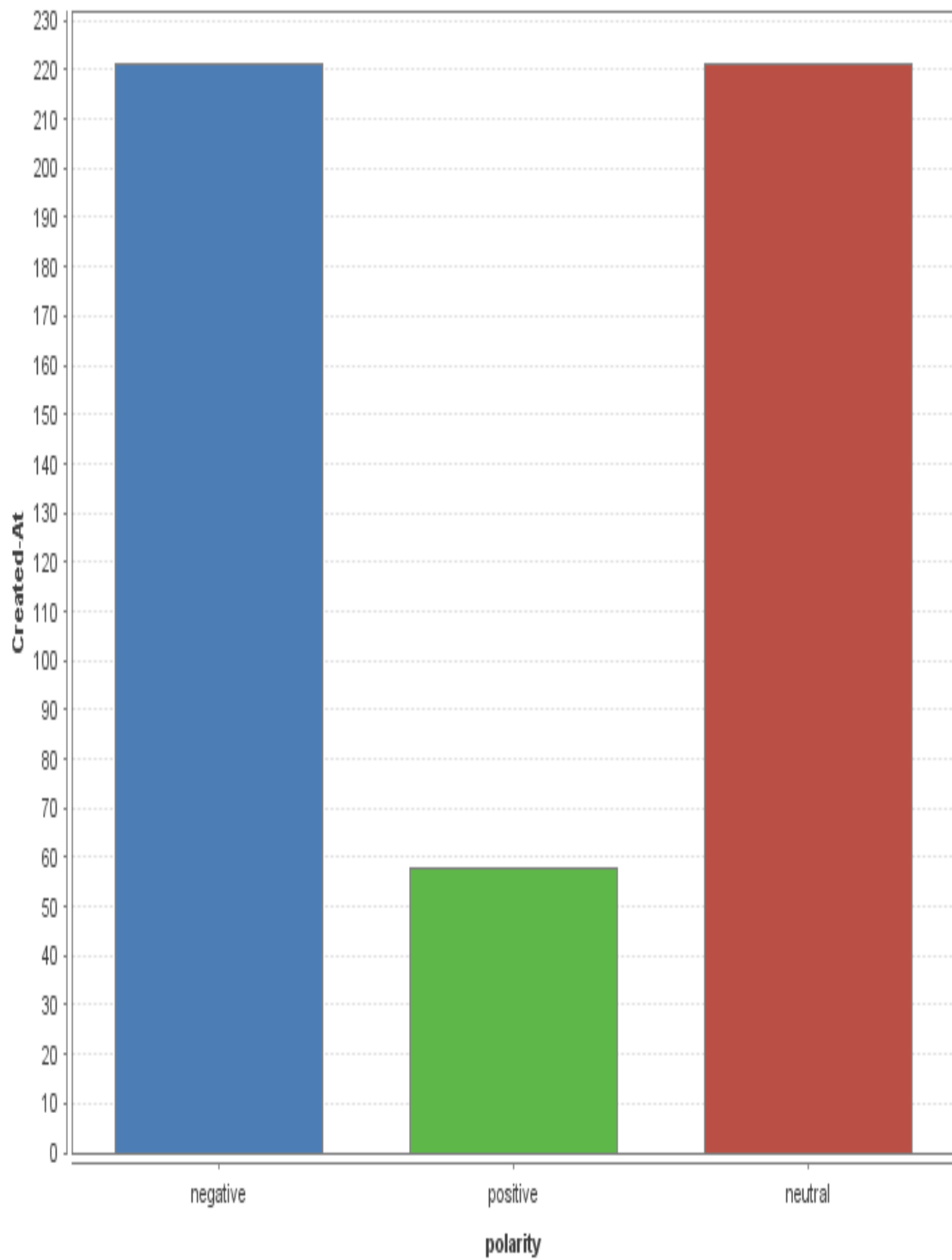
Donald Trump tweets:



**Fig 9**

From the figure (Fig 9), the most tweets are neutral or negative on Donald Trump and only a few tweets are positive. This provides an idea on an individual's (in twitter) opinion on this specific person.

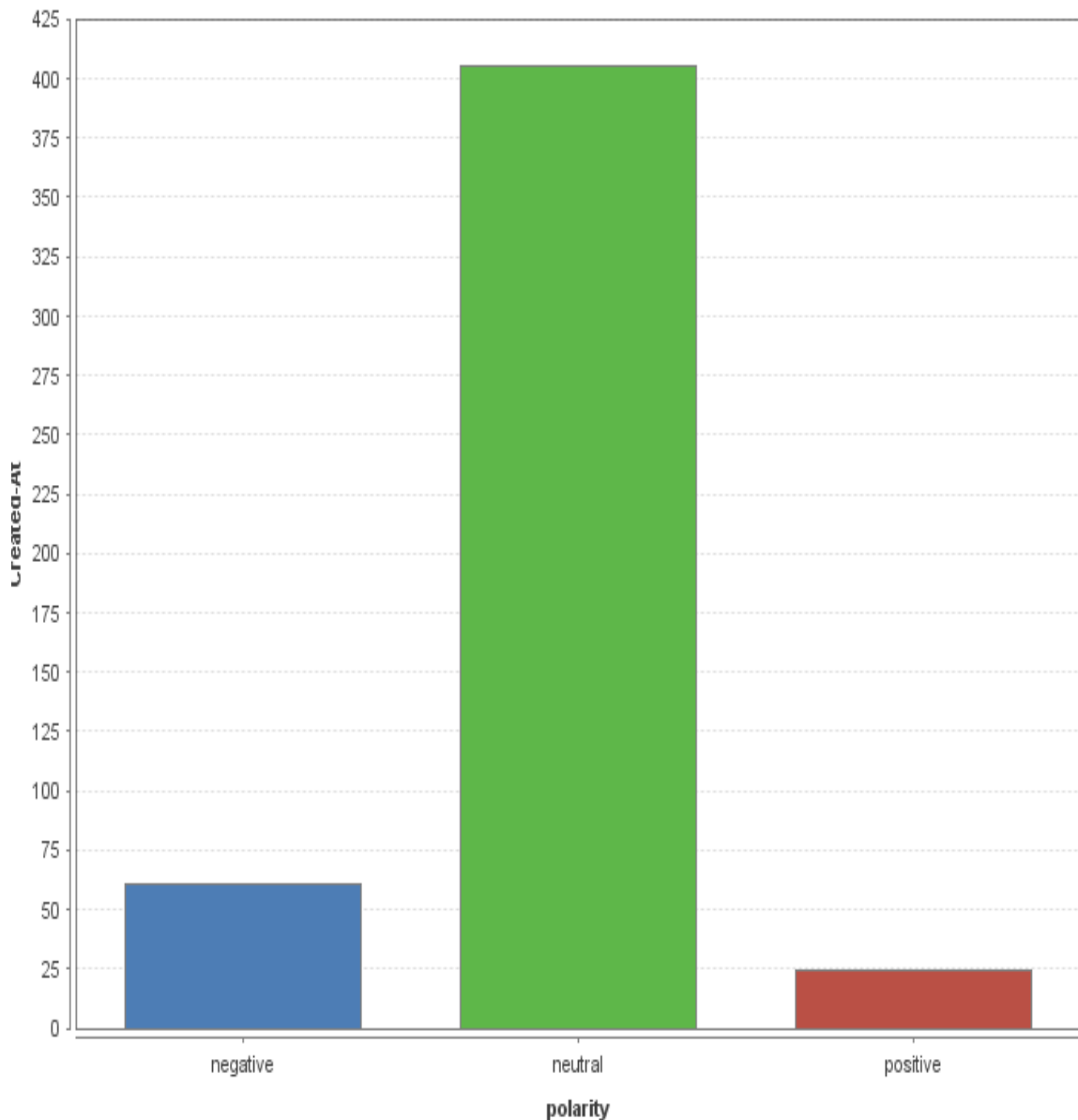
Barack Obama tweets:



**Fig 10**

The above plot (Fig 10) shows that Barack Obama has equal number of negative and positive tweets on him.

Narendra Modi tweets:



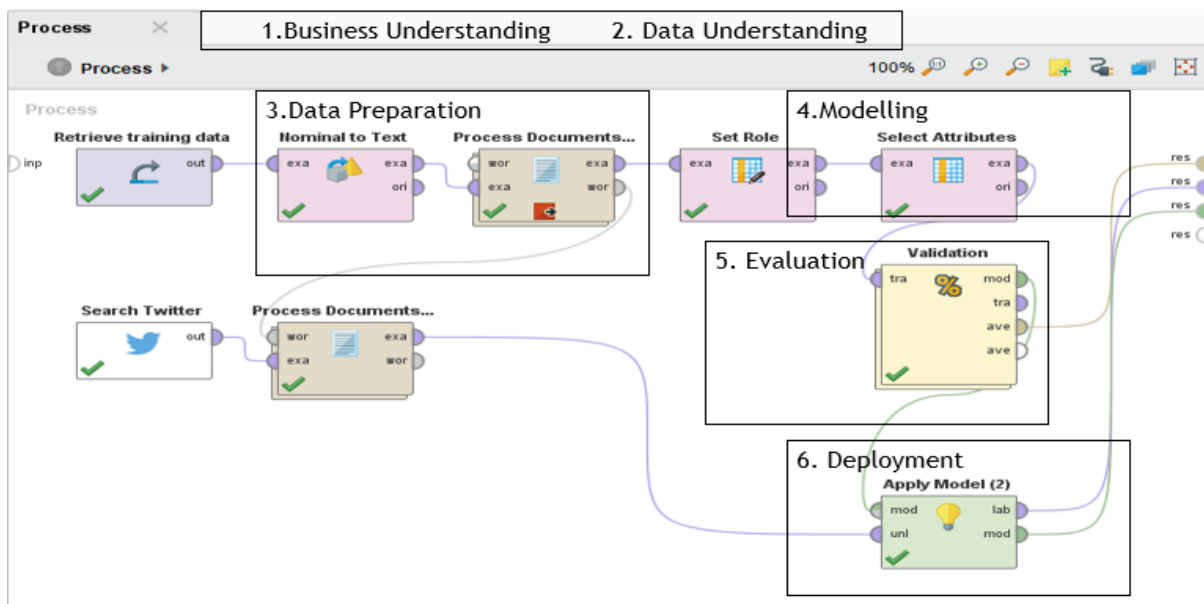
**Fig 11**

The obtained figure (Fig 11) shows that the tweets on Narendra Modi are mostly neutral and the number of positive and negative tweets are comparatively equal.

The results suggest that model is effective in analysing the opinion and sentiment of people from twitter. Given a specific topic, the model can accurately identify the sentiment towards the topic from a target audience. This generally addresses the overall business problem stated in the proposal of the project.

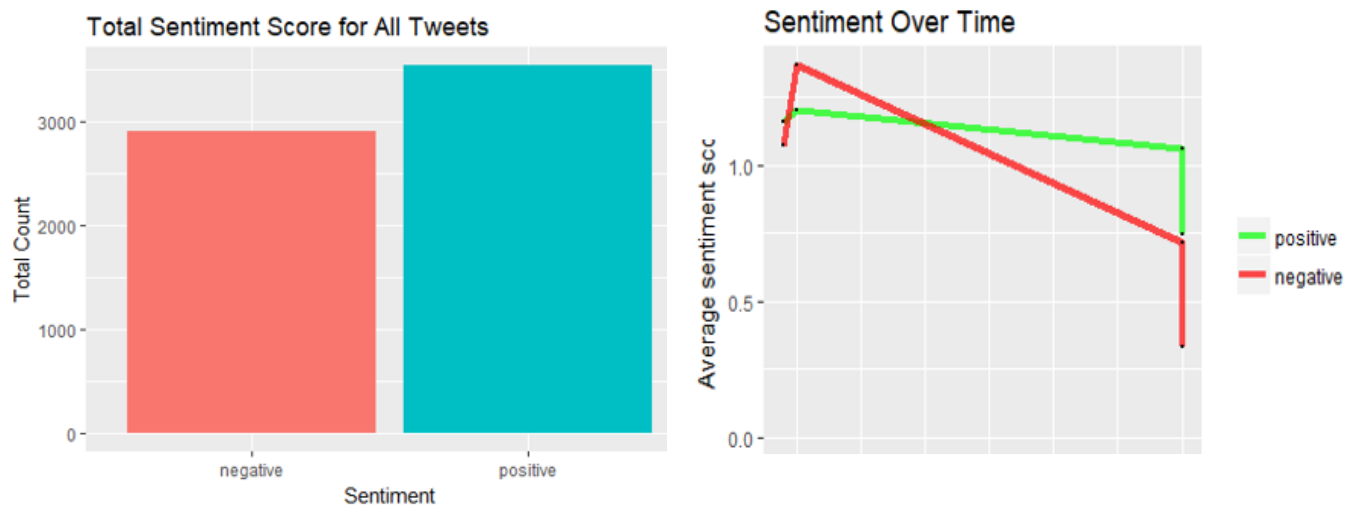
# Conclusion

The project successfully followed the CRISP-DM methodology to reach the suggested goal. Figure 12 shows the whole designed process in RapidMiner along with the labelled boxes that show the different phases of the CRISP-DM methodology.



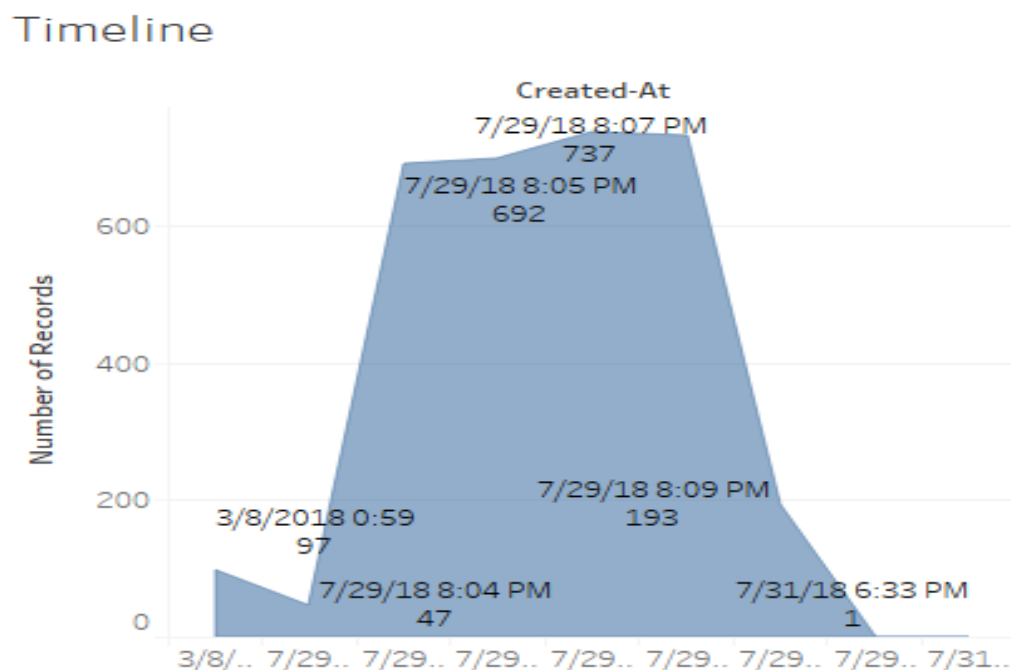
**Fig 12**

The first two phases of CRISP-DM methodology occur before the design as these phases involve understanding the business and the data which here is understanding the twitter platform, sentiment analysis methods, the applications of sentiment analysis and understanding the dataset. The rest of the phases are in the design which involves modelling, validation and deployment. Displaying the result of the project by successfully analysing and visualizing the polarity of tweets on leaders of different countries is considered as the deployment phase. The results from these datasets can further be used to visualize the polarities on other data visualizing platforms. The output figures (Fig 13 &14) are example visualizations from Tableau and R (using ggplot2).



**Fig 13**

Figure 13 shows the total polarity strength of and the variation with time tweeted on Narendra Modi tweets visualized in R. Figure 14 shows the trend in tweeting time of people from the Trump dataset visualized using Tableau.



**Fig 14**

To conclude, the sentiment analysis model has been successfully created and the model was used to explore and analyse the sentiment of the stated business problem



# **References**

Kaggle.com. (2018). *Kaggle: Your Home for Data Science*. [online] Available at: <https://www.kaggle.com/> [Accessed 23 Jul. 2018].

RapidMiner. (2018). *Visual Workflow Designer for Data Scientists / RapidMiner Studio*. [online] Available at: <https://rapidminer.com/products/studio/> [Accessed 17 Jul. 2018].

Rdocumentation.org. (2018). *R Documentation and manuals / R Documentation*. [online] Available at: <https://www.rdocumentation.org/> [Accessed 15 Jul. 2018].

Tableau Software. (2018). *Tableau Software*. [online] Available at: <https://www.tableau.com/> [Accessed 1 Aug. 2018].

Wikipedia.org. (2018). *Wikipedia*. [online] Available at: <https://www.wikipedia.org/> [Accessed 4 Aug. 2018].