Faculty of Engineering and Technology
Project Work – Student Log Book

Degree/ program		Specialisation	Project Team ID
B.Tech		Computer Science Engineering	J320

Name of student	Register Number	Department	Mobile Number	Email ID
Mayank Bansal	RA141100301694	CSE	9176283774	bansal1996@outlook.com
Ujjwal jain	RA141100301646	CSE	9962727940	ujjwaljain423@gmail.com

Academic Year	2017	Semester	8
Course Code	CS-1050	Course Title	Major Project

Working Title of the Project:	**SENTIMENT ANALYSIS ON TWITTER DATA**
Project Site / Location	SRM University
Name and address of the company / organisation (Applicable for projects with industry or industry support)	SRM institute of technology and science

Supervision Team

	Supervisor	**Co-Supervisor**	**External Supervisor (If applicable)**
Name	Saad Yunus Sait		
Designation	Research Associate Professor		
Department	CS&E		
Campus	KTR		
Telephone	9884355338		
E-mail	saad.y@ktr.srmuniv.ac.in		

Mission Statement

Problem (or) Product Description
Product Description
Sentiment analysis has become a buzzword lately as social networks are bustling with consumer chatter. Sentiment analysis, also known as opinion mining, is the application of Natural Language Processing (NLP) techniques and text analytics for identifying patterns and extracting insights from consumer data. It is widely used by brands to gauge the reactions of their consumers towards their products, product features, promotional campaigns and so on. The problem in sentiment analysis is classifying the polarity of a given text at the document sentence, or feature/aspect level . Whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral.
Assumptions and Constraints
Only english language can be analysed for sentiment analyse.
Stakeholders
Mayank Bansal
Ujjwal Jain

Faculty of Engineering and Technology

Project Work – Student Log Book

SRM UNIVERSITY (Under section 3 of UGC Act 1956)

Division of work and contributors

Time period		Activities or components of the project	Name/Register Number of the Individual Contributor	Names/Register Number of the Joint Contributors
From Date	**To Date**			
1/11/2017	8/11/2017	initial setup of development environment for R language which involved downloading r studio. and getting access from twitter for fetching tweets.	RA1411003010694	RA1411003010646
10/11/2017	17/11/2017	explored different possible algorithm for text analysis and how can they be used to improve the current methods of sentiment analysis. Initial pseudo code prepared and run on tweets to fetch them and pre-process them for further analysis. Word database of positive and negative tweets was downloaded and initial comparison of tweets to classify the according to number of positive words and negative.	RA1411003010694	RA1411003010646
19/11/2017	25/11/2017	First result was displayed using using different charts format like PI chart and histogram. The initial test is done on #arunavmodi scam which showed how people are reacting about the scam. the result shown that mostly people are being negative about the scam.	RA1411003010694	RA1411003010646
5/1/2018	14/1/2018	changes are made in the plot function to show labels and other separator.	RA1411003010646	
18/1/2018	24/1/2018	Machine learning algorithm is explored to incorporate natural language processing to get more accurate result.Logistic regression model is used for classifying tweets based on probability. initial data set of 1.6million tweets is taken from analyticsvidhya for training the model and the whole data set is divided into two parts one for training and other for testing the model. R library doc2vec is used to genrate document term matrix from the given corpus of training data set	RA1411003010646	RA1411003010694
25/1/2018	2/2/2018	The tweets are pre-processed in order to remove unwanted data which does not contribute to overall sentiment, like url and other words. the initial tokenizer is made with the help of itoken function provided in doc2vec library	RA1411003010646	RA1411003010694
3/1/2018	9/2/2018	function to change the sentiment values in the given data set is created, it help to classify the data into two binary variable 0 and 1 where 1 being positive and 0 being negative	RA1411003010694	RA1411003010646

10/2/2018	15/2/2018	function to partition the data set for training and testing, in which 80% of the data is used for training and 20% is used for testing.	RA1411003010694	
18/2/2018	24/2/2018	Function to train the model using the glmnet package is written and first model is generated.	RA1411003010646	
25/2/2018	1/3/2018	explored different methods to overcome the overfitting issue in the training of the model.	RA1411003010694	
3/3/2018	10/3/2018	program to fetch the tweet from twitter and process them to generate the input matrix to predict with model generated.	RA1411003010646	
11/3/2018	19/3/2018	program to plot the graph using ggplot library in R. and different possble ways the graph can be made more visually understandable	RA1411003010646	
22/3/2018	28/3/2018	function to plot the graph is modified to pass different color scheme and using different separator for the graph.	RA1411003010694	
1/4/2018	7/4/2018	introduction chapter is written for the report and other section is explained in detail in the report.	RA1411003010694	
7/4/2018	12/4/2018	coding and implementation chapter and chapter about sentiment analysis with different approach is written.	RA1411003010646	

Summary record of major progress meetings with supervisors

Summary record of major progress meetings with supervisors			**Working title of dissertation/research project:**	
Meeting date & supervisors present	**Progress since last meeting**	**Agreed programme of work and target dates**	**Other issues, e.g. facilities, supervision, training needs, etc.**	**Date of next meeting**
8/11/2017	in the first meeting , the different methods of sentiment analysis was explored and knowledge based approach was decided to work with.			14/11/2017

14/11/2017	we used the knowledge based approach to classify the tweets according to sentiment score being positive or negative.	The bigger words database should be used for analyzing the tweets according to positive database and negative word database. TD: 22/11/2017		5/1/2018
5/1/2018	we used bigger dataset for positive words and negative words and the result was more accurate for the same test that we did earlier with small words database.	we decided to incorporate the emoticon in the sentiment to get the more accurate result.		11/1/2018

Summary record of major progress meetings with supervisors

Summary record of major progress meetings with supervisors			**Working title of dissertation/research project:**	
Meeting date & supervisors present	**Progress since last meeting**	**Agreed programme of work and target dates**	**Other issues, e.g. facilities, supervision, training needs, etc.**	**Date of next meeting**
11/1/2018	we made wordcloud to for visual analysis of the whole experiment . The first test was done on arunav modi scam.	Try to implement machine learning in the analysis and use natural language processing to get more accurate result instead of knowledge based approach. 18/1/2018		20/1/2018
20/1/2018	this is the first review with teacher and first demo of the working model.	Use the logistic regression model to classify the tweets and use a bigger dataset for training the model. 29/1/2018		
29/1/2018	The first data set of the 1.6million is taken and we start partitioning the dataset in two parts for training the model and for testing the model.	pre-process the data by using tokenizer function in R, and use vocabulary approach to generate a document term matrix. 8/2/2018	The data set had tweets classified positive as 4 and negative as 0, and our requirement is to mutate them like, positive as 1 and negative as 0.	9/2/2018

Summary record of major progress meetings with supervisors

Summary record of major progress meetings with supervisors			Working title of dissertation/research project:	
Meeting date & supervisors present	**Progress since last meeting**	**Agreed programme of work and target dates**	**Other issues, e.g. facilities, supervision, training needs, etc.**	**Date of next meeting**
9/2/2018	Dataset is mutated and tweets were classified , positive as 1 and negative as 0. some tweets in the dataset had NA, which means there was no sentiment mentioned so that was removed to clean the dataset.	use 20% of the data for testing and 80% for training the model, and use glmnet library to train the model. 15/2/2018		16/2/2018
16/2/2018	first model is prepared using the training dataset, with 0.875 area under the curve of the logistic function.	how to avoid overfitting of the model by k-fold cross validation to make the model more generalized, so that it will work better on unknown data set. 22/2/2018		27/2/2018
27/2/2018	Second model was prepared by using the k-fold cross validation to overcome the overfitting of model, and prediction was performed on testing dataset.	Save the trained model for future use as it takes so much time to train the model. and use ggplot library to display the graph. 5/3/2018		6/3/2018
6/3/2018	The accuracy of the model is checked by doing prediction against the data fetched from twitter and it also pre-processed the same as our training dataset.	use more number of tweets as twitter only allows to fetch 250 tweets in free api account, and how the graph should look like to display it more visually understandable. 14/3/2018		14/3/2018

14/3/2018	we performed different experiments on many topics to check the working and accuracy of the model by taking different input each time for each topic.	try to use the separator for showing tweets as negative and positive in the graph and use different color scheme. 20/3/2018		20/3/2018
20/3/2018	current graph looks more understandable in terms of classifying and seeing the difference of each tweet and its probability.	Use legends in the graph to denote what color is used for which clssification like positive and negative and neutral. 27/3/2018		27/3/2018
27/3/2018	The project is succfully implemented and different test is performed and the graph looks even more understandable from the last meeting.	start making the report and describe each section of the project and mentioned both approach of the sentiment analysis and show a comparison between them. 4/4/2018		4/4/2018
4/4/2018	50% report is prepared and new section is introduced in which we wrote about the previous work done in the field of sentiment analys and their contribution.	Use different figure to explain the difference of the approach and their use cases, use mathematical formula to explain the models		11/4/2018
11/4/2018	report is almost complete and some minor changes nees to be made.	prepare the presentation for the final review and get the plagiarism check done for the report. 18/4/2018		18/4/2018

Worksheet / Data collection / Observation etc

1. Initial word database of positive words and negative words is collected.
2. Result about arunavmodi scam is observed and sentiment score is negative
3. dataset of 1.6million tweets is collected from analyticsvidhya website.

Worksheet / Data collection / Observation etc

4. area under the curve on training data set is 0.875
5. Prediction is accurate as auc on testing data is 0.874

Faculty of Engineering and Technology

Project Work – Student Log Book

Project Assessment (Reviews) Date:

Zeroth Review

Register Number	General comments	Specific comments	Title of the Project	Total 15 Marks
RA1411003010694				
RA1411003010646				

Project Assessment (Reviews) Date:

Review: I

Register Number:

Faculty of Engineering and Technology

Project Work – Student Log Book

General comments	Specific comments	Reviewer Names	Literature Survey	Project Demo	Architecture Diagram	Explanation/Algorithm used	Total
			10 Marks	10 Marks	5 Marks	5 Marks	30 marks

IMPLEMENTATION: 40% OF CODE

Project Assessment (Reviews) Date:

Review: I

Register Number:

Faculty of Engineering and Technology

Project Work – Student Log Book

General comments	Specific comments	Reviewer Names	Literature Survey	Project Demo	Architecture Diagram	Explanation/Algorithm used	Total
			10 Marks	10 Marks	5 Marks	5 Marks	30 marks

IMPLEMENTATION: 20% OF CODE

Faculty of Engineering and Technology

Project Work – Student Log Book

Project Assessment (Reviews) Date:

Review: II

Register Number:

General comments	Specific comments	Reviewer Names	Presentation	Design & Methodology	Implementation	Total
			10 Marks	20 Marks	20 Marks	50 Marks

IMPLEMENTATION: 70% OF CODE

Faculty of Engineering and Technology

Project Work – Student Log Book

Project Assessment (Reviews) Date:

Review: II

Register Number:

General comments	Specific comments	Reviewer Names	Presentation	Design & Methodology	Implementation	Total
			10 Marks	20 Marks	20 Marks	50 Marks

IMPLEMENTATION: 70% OF CODE

Project Assessment (Reviews) Date:

Faculty of Engineering and Technology

Project Work – Student Log Book

SRM UNIVERSITY
(Under section 3 of UGC Act 1956)

Review: III

Register Number:

General comments	Specific comments	Reviewer Names	Presentation	Demonstration	Report	Performance matrices and result	Journal Publication Acceptance	Total
			10 Marks	25 Marks	10 Marks	10 Marks	30 Marks	105 Marks

IMPLEMENTATION: 100% OF CODE

Faculty of Engineering and Technology

Project Work – Student Log Book

Project Assessment (Reviews) Date:

Review: III

Register Number:

General comments	Specific comments	Reviewer Names	Presentation	Demonstration	Report	Performance matrices and result	Journal Publication Acceptance	Total
			10 Marks	25 Marks	10 Marks	10 Marks	30 Marks	105 Marks

IMPLEMENTATION: 100% OF CODE