|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Working Title of the Project:** | | **Opinion Mining in Twitter using Machine Learning** | | |
| **Project Site / Location** | | **CHENNAI** | | |
| **Name and address of the company / organisation**  **(Applicable for projects with industry or industry support)** | | **SRM University, Kattankulathur, Kanchipuram district - 603203** | | |
| **Supervision Team** | | | | |
|  | **Supervisor** | | **Co-Supervisor** | **External Supervisor**  **(If applicable)** |
| **Name** | **S. Thenmalar** | |  |  |
| **Designation** | **Assistant Professor** | |  |  |
| **Department** | **CSE** | |  |  |
| **Campus** | **Kattankulathur** | |  |  |
| **Telephone** | **9841244268** | |  |  |
| **E-mail** | **thenmalar.s@ktr.srmuniv.ac.in** | |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name of student | Register Number | Department | Mobile Number | Email ID |
| Sharad Kakran | 1031310379 | CSE | 9962644935 | sharadkakran@gmail.com |
| Shantanu Garg | 1031310362 | CSE | 8754512877 | shantanu.garg@ymail.com |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Degree/ program | B.Tech | Specialisation | Computer Science Engineering  **Project Team ID** |
| Academic Year | 2017 | Semester | 8 |
| Course Code | CS1050 | Course Title | Major Project |
| Mission Statement | | | | |
| **Problem (or) Product Description** | | | | |
| In today’s world, where the outreach of internet is rapidly increasing, The internet has empowered people in a lot of ways. The internet has become the prime source of information for everyone today. | | | | |
| People take on the internet to express their opinions on various topics ranging from the political party they prefer to the brand of ramen they like. | | | | |
| It is very important for brands & companies to keep a track of the public sentiment online in order to serve their customers better. | | | | |
| Sentiment analysis in twitter using machine learning techniques presents us with a way of classifying tweets posted by users on twitter in various emotions, thus helping us better understand the mood of the customer. | | | | |
| Algorithms such as LSA, NB, SVM and KNN have been used in this project in order to classify and understand the underlying emotions of the tweets. | | | | |
|  | | | | |
| **Assumptions and Constraints** | | | | |
| There is a lot of noise present in twitter data. Special symbols, smileys etc. are some of the things which constitute the noise. To remove the noise from the tweets, the data has been cleaned in pre-processing. | | | | |
| Unstructured data slangs such as LOL ( Laugh out loud) and OMG ( Oh My God) can cause some problems and result in misclassification. | | | | |
| Ambiguous words can often change the meaning of the sentence. Presence of ambiguous words is an issue that is still challenging to deal with. | | | | |
| Sarcasm is difficult to deal with, when present in a dataset. It increases the ambiguity of the sentence and often results in misclassification of a sentence. | | | | |
|  | | | | |

**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** |
| 01/01/17 | 01/01/17 | Setting up Linux Environment in virtual box | 1031310379 |  |
| 02/01/17 | 03/01/17 | Simulation of the Linux environment | 1031310362 |  |
| 04/01/17 | 06/01/17 | Performing test runs |  | 1031310362  1031310379 |
| 08/01/17 | 09/01/17 | Installation of required repositories and kernels in the Linux environment | 1031310362 |  |
| 12/01/17 | 15/01/17 | Extraction of data from twitter using API |  | 1031310362  1031310379 |
| 17/01/17 | 18/01/17 | Segregation of tweets | 1031310379 |  |
| 20/01/17 | 21/01/17 | Classification of tweets in Context-Specific category | 1031310379 |  |
| 23/01/17 | 25/01/17 | Classification of tweets in Random category |  | 1031310362  1031310379 |
| 28/01/17 | 29/01/17 | Removal of misspelling and slang words from the dataset | 1031310362 |  |
| 30/01/17 | 31/01/17 | Removal of smileys and emoticons from the dataset | 1031310362 |  |
| 03/02/17 | 04/02/17 | Removal of punctuation marks and other noises from the data set | 1031310379 |  |
| 06/02/17 | 08/02/17 | Division of dataset into Test data and Training data |  | 1031310362  1031310379 |
| 10/02/17 | 11/02/17 | Improvising Feature extraction | 1031310362 |  |
| 12/02/17 | 15/02/17 | Executing bag of words technique on the dataset |  | 1031310362  1031310379 |
| 16/02/17 | 16/02/17 | Tagging the data with different parts of speech | 1031310379 |  |

**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** |
| 17/02/17 | 19/02/17 | Implementation of Bi grams | 1031310379 |  |
| 21/02/17 | 24/02/17 | Implementation of Trigrams |  | 1031310362  1031310379 |
| 25/02/17 | 27/02/17 | Reduction of the size of dataset using chi square test |  | 1031310362  1031310379 |
| 28/02/17 | 28/02/17 | Testing and Result Analysis | 1031310362 |  |
| 01/03/17 | 03/03/17 | Implementation of LSA algorithm on random tweets |  | 1031310362  1031310379 |
| 04/03/17 | 04/03/17 | Implementation of KNN algorithm on Random tweets | 1031310379 |  |
| 08/03/17 | 10/03/17 | Implementation of SVM algorithm on Random tweets |  | 1031310362  1031310379 |
| 11/03/17 | 11/03/17 | Implementation of NB algorithm on random tweets | 1031310362 |  |
| 12/03/17 | 12/03/17 | Classification based on moods | 1031310362 |  |
| 13/03/17 | 15/03/17 | Results analysis |  | 1031310362  1031310379 |
| 16/03/17 | 18/03/17 | Implementation of LSA algorithm on context specific tweets |  | 1031310362  1031310379 |
| 19/03/17 | 20/03/17 | Implementation of KNN algorithm on context specific tweets | 1031310379 |  |
| 22/03/17 | 24/03/17 | Implementation of SVM algorithm on context specific tweets |  | 1031310362  1031310379 |
| 26/03/17 | 28/03/17 | Implementation of NB algorithm on context specific tweets |  | 1031310362  1031310379 |
| 29/03/17 | 31/03/17 | Classification based on moods |  | 1031310362  1031310379 |

**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** |
| 02/04/17 | 02/04/17 | Results analysis | 1031310379 |  |
| 03/04/17 | 03/04/17 | Tweaking of variables | 1031310379 |  |
| 04/04/17 | 05/04/17 | Test run and accuracy analysis | 1031310362 |  |
| 06/04/17 | 07/04/17 | Computation of the impact of two approaches |  | 1031310362  1031310379 |
| 08/04/17 | 08/04/17 | Improving the accuracy of the model | 1031310379 |  |
| 09/04/17 | 10/04/17 | Plotting of graphs | 1031310379 |  |
| 12/04/17 | Ongoing | Accuracy analysis |  | 1031310362  1031310379 |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**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** |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**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** |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**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** |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**Worksheet / Data collection / Observation etc**

A.) Latent Semantic Analysis (LSA)

Latent Semantic Analysis is also known as Latent Semantic Index (LSI). It is used extensively in text-based sentiment

classification. Its purpose is to analyse text documents and understand the latent meaning of those documents. The LSA

algorithm follows a two-point approach in order to classify a document. The first step is called the bag of words approach.

Here, the documents are represented as “bag of words”. A count matrix is prepared where the number of instances of each word

is calculated. In this step, the order of words is not important. In the second step, the algorithm tries to look for patterns for

words that usually occur together in a certain kind of document. Next, we try to identify the frequency of a word in each

document following which we can assign weights to each word.

The most popular weighting technique is called TFIDF.

TFIDFi,j = ( Ni,j / N\*,j ) \* log( D / Di )

In this formula, TFIDF is the term frequency in the document, Ni,j represents the number of times the word i appears in the

document j, N\*,j represents the number of total words in the document j, D is the number of documents and Di is the number

of documents in which word i appears.

B.) Naïve Bayes Algorithm (NB)

Naïve Bayes is a collection of classification algorithms which are based on the Bayes theorem. The common principle that

each algorithm in the Naïve Bayes family shares is that every feature being classified is independent of the value of any other

feature. Naïve Bayes is a simple algorithm than can sometimes outperform more sophisticated algorithms.

The conditional probability for the Naïve Bayes is defined as :

P( Xj yj) = \_mi=1 P( xij yj) (2)

In this formula, is the feature vector defined as X=fx1 ,x2 ,....xm g and yj is the class label.

C.) Support Vector Machines (SVM)

Support Vector Machines aka SVM is a machine learning problem that is widely used in classification problems. SVM’s

are based on the idea of finding the right hyperplane that can divide a dataset into two classes. SVM uses the discriminative

**Worksheet / Data collection / Observation etc**

function defined as

g( X) = wT \_( X) + b (3)

In this formula, ’X’ is the feature vector, ’w’ is the weights vector and ’b’ is the bias vector. \_ () is the non-linear mapping

from input space to high dimensional feature space. ’w’ and ’b’ are learned automatically on the training set.

D.) K-Nearest Neighbours (KNN)

K-Nearest Neighbours algorithm or the KNN algorithm is a non-parametric lazy learning algorithm. This algorithm stores

all available cases and classifies new cases based on similarity measure. It measures the distance between a query scenario and

a set of other scenarios in the dataset. KNN algorithm is measured by a distance function.

dist(x, y) = n i=1Σ √(xi ). - (yi). (4)

In this formula, We can compute the distance between two

scenarios using some distance function d(x,y), where x,y are

scenarios composed of N features.

Project Assessment (Reviews) Date: 20/10/16

**Zeroth Review**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Register Number** | **General comments** | **Specific comments** | **Title of the Project** | **Total** |
| **20 Marks** |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Project Assessment (Reviews) Date: 27/01/17

**Review: I**

**Register Number: 1031310379**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **General comments** | **Specific comments** | **Reviewer Names** | **Presentation** | **Design &**  **Methodology** | **Implementation** | **Report** | **Total** |
| **10 Marks** | **15 Marks** | **15 Marks** | **10 Marks** | **50 marks** |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

IMPLEMENTATION: 20% OF CODE

Project Assessment (Reviews) Date:

**Review: I**

**Register Number: 1031310362**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **General comments** | **Specific comments** | **Reviewer Names** | **Presentation** | **Design &**  **Methodology** | **Implementation** | **Report** | **Total** |
| **10 Marks** | **15 Marks** | **15 Marks** | **10 Marks** | **50 marks** |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

IMPLEMENTATION: 20% OF CODE

Project Assessment (Reviews) Date:

**Review: II**

**Register Number: 1031310379**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **General comments** | **Specific comments** | **Reviewer Names** | **Presentation** | **Design &**  **Methodology** | **Implementation** | **Report** | **Total** |
| **10 Marks** | **15 Marks** | **15 Marks** | **10 Marks** | **50 Marks** |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

IMPLEMENTATION: 50% OF CODE

Project Assessment (Reviews) Date:

**Review: II**

**Register Number: 1031310362**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **General comments** | **Specific comments** | **Reviewer Names** | **Presentation** | **Design &**  **Methodology** | **Implementation** | **Report** | **Total** |
| **10 Marks** | **15 Marks** | **15 Marks** | **10 Marks** | **50 Marks** |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

IMPLEMENTATION: 50% OF CODE

Project Assessment (Reviews) Date:

**Review: III**

**Register Number: 1031310379**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **General comments** | **Specific comments** | **Reviewer Names** | **Presentation** | **Design &**  **Methodology** | **Implementation** | **Report** | **Journal Publication** | **Total** |
| **10 Marks** | **15 Marks** | **15 Marks** | **10 Marks** | **30 Marks** | **80**  **Marks** |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

IMPLEMENTATION: 100% OF CODE

Project Assessment (Reviews) Date:

**Review: III**

**Register Number: 1031310362**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **General comments** | **Specific comments** | **Reviewer Names** | **Presentation** | **Design &**  **Methodology** | **Implementation** | **Report** | **Journal Publication** | **Total** |
| **10 Marks** | **15 Marks** | **15 Marks** | **10 Marks** | **30 Marks** | **80**  **Marks** |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

IMPLEMENTATION: 100% OF CODE