

Software Project Management Report on

**Mining Indian Tweets to Understand
Food Price Crises**



National Institute of Technology Karnataka Surathkal
October 26, 2017

Under the guidance of

Dr. Mahendra Pratap Singh

Submitted by

Bairi Sandhya Rani	14C0205
Dokku Tejaswi Surya Sai Gayathri	14C0211
Moore Saranya	14C0226

CONTENTS

1. Objective	1
2. Project Description	1
2.1. Requirements	2
2.2. Challenges	2
3. GANTT Chart	2
4. PERT Chart	8
5. WBS	9
6. CPM Chart	10
7. Team Skills	12
8. Conclusion	13

1. Objective

The main objective of this project is to monitor the Twitter conversations in India to understand how the conversation volume trends, variation of the sentiment and opinions relate to actual events specifically related to food price. We are operating on the premise that online social media conversations might represent a new source of information.

2. Project description

The globalization of world economy requires the basic statistics on supply and demand for agriculture. The rising food prices and food shortage highlights the need for more concrete efforts to harness the potential of the agricultural sector for development since it has remained the dominant sector in India's economy. Traditional statistics, household surveys and census data have been effective in tracking medium to long-term development trends, but are less effective in generating a real-time snapshot in order for policymakers to develop timely actions to protect vulnerable populations against crises. Better data and statistics will help governments to track progress and ensure their decisions are evidence-based.

Since millions of users share their opinions on twitter, making it a valuable platform for tracking and analyzing public opinion, it can provide critical information for us in various domains. This analysis includes mainly tweets collection based on taxonomy that consists of words and phrases related to prices of food, using either REST or Streaming APIs in python language. The collected data needs to be preprocessed which involves tokenization, filtering (removing hashtags, RT, @), stemming etc. The pre-processed dataset has various discrete properties. It is exposed to feature extraction methods like terms frequency, negative phrases, parts of speech etc., We extract different aspects which are later identified as positive, negative or neutral to detect the polarity of the whole sentence. This comes under Sentiment Classification for which machine learning techniques like Support Vector Machine (SVM), Naive Bayes Classifier, Maximum Entropy, Decision Tree etc. can be used.

All the relevant data is analyzed to provide a proportion of tweets related to each theme to determine the statistical pattern of conversation for each category. The general volume of relevant tweets, independent from the conversation is also to be analyzed to find relations with official statistics.

Automated monitoring of public sentiment on social media, combined with contextual knowledge, has the potential to be a valuable real-time proxy for food-related economic indicators. Current challenges to overcome include establishing high frequency models of food prices and validating them using official statistics, filtering out noise due to non-relevant news items etc. If social media data mining to model food prices matures to become robust in the future, statistical institutes might consider including social media monitoring into official statistics channels.

2.1 Requirements

Functional requirements - They define what a software project is supposed to do. Extraction of relevant tweets, preprocessing, evaluation of sentiment of tweets and predicting food crisis in the future based on this analysis are the key functional requirements for this project.

Non-functional requirements - They define how a software project is supposed to be. Accessibility, compatibility, portability, reliability and testability are some of the non functional requirements to be taken into consideration for this project.

2.2 Challenges


The collection of enormous amount of data on twitter is one of the biggest challenges of this project. It took almost 10 days for reliable data collection.

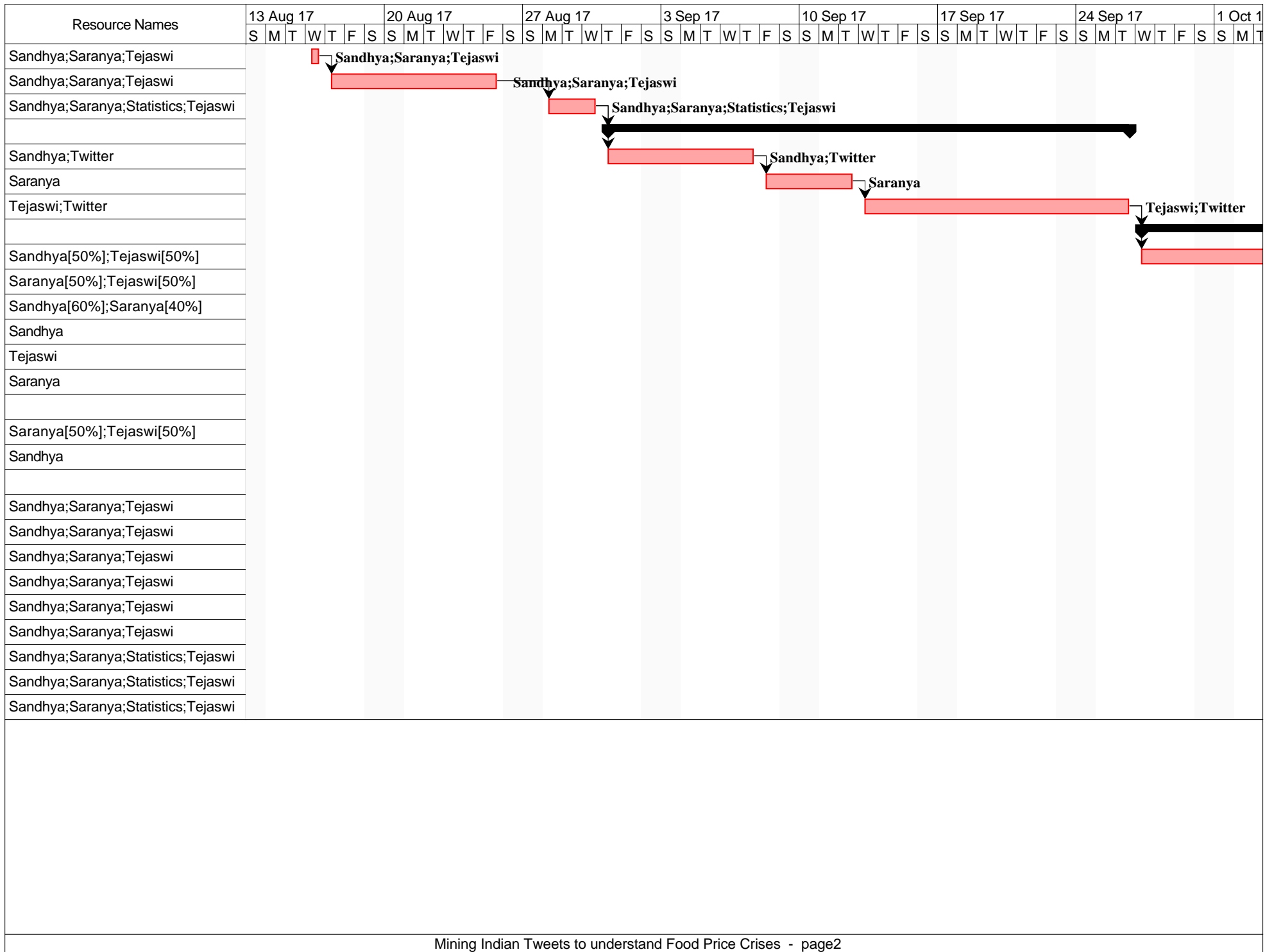
The second challenge is the pre-processing the raw data applying classifiers. A step by step procedure is followed to complete this task. It took us almost a month to get a satisfactory output and we are still working for better results.

Identifying suitable feature extraction method and building and training the classifier using training data are the other important phases where there is high possibility of facing difficulties.

3. Gantt chart

It is a type of bar chart which visually represents large projects broken into smaller tasks or activities with each of these activities spread over a period of time. It illustrates project schedule and the start and finish dates of the summary and terminal elements of the project. The following gantt chart has been created using Projectlibre providing sufficient details of tasks, start and finish dates.

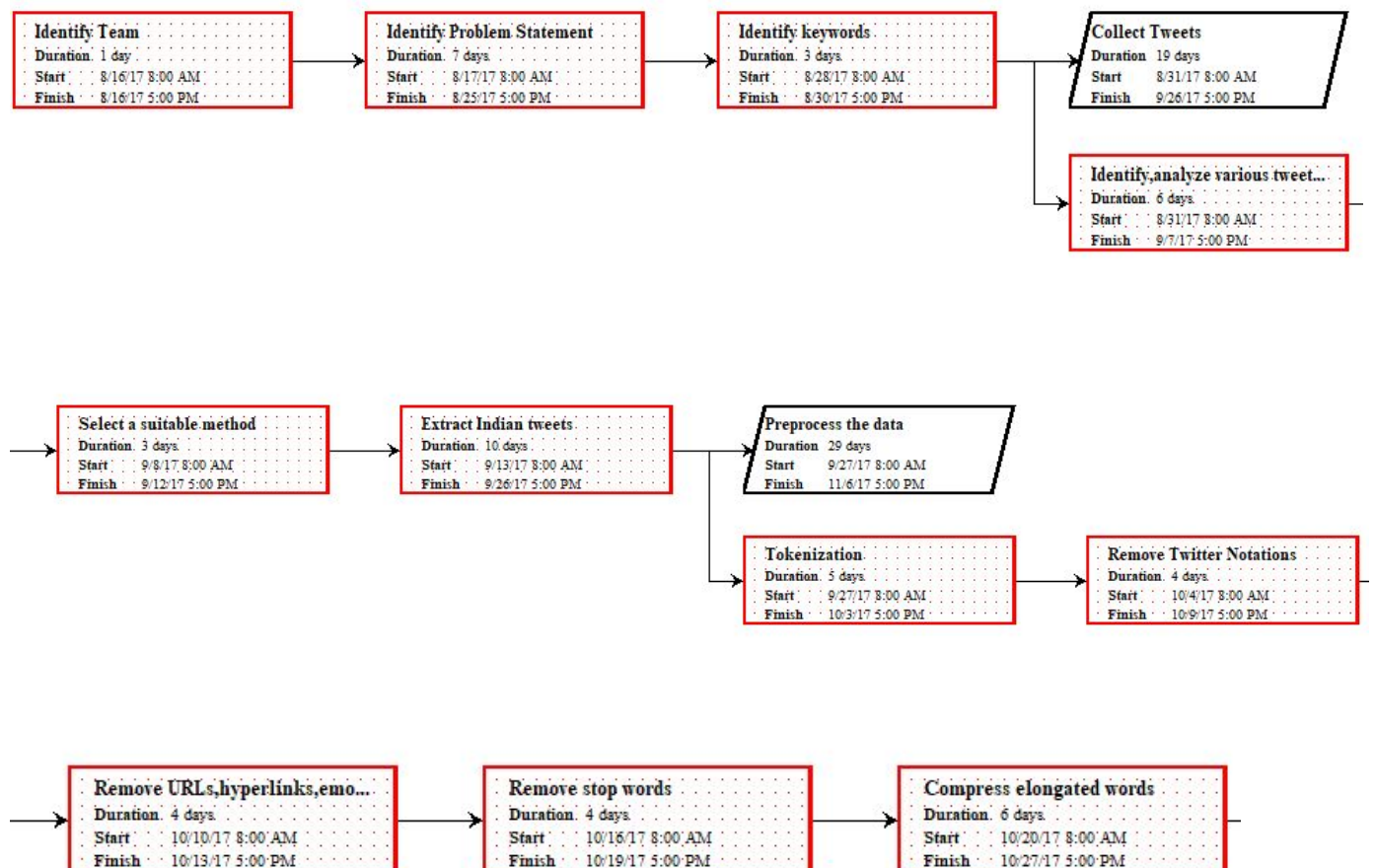
		Name	Duration	Start	Finish	Predecessors
1		Identify Team	1 day	8/16/17 8:00 AM	8/16/17 5:00 PM	
2		Identify Problem Statement	7 days	8/17/17 8:00 AM	8/25/17 5:00 PM	1
3		Identify keywords	3 days	8/28/17 8:00 AM	8/30/17 5:00 PM	2
4		Collect Tweets	19 days	8/31/17 8:00 AM	9/26/17 5:00 PM	3
5		Identify,analyze various tweet collection methods	6 days	8/31/17 8:00 AM	9/7/17 5:00 PM	3
6		Select a suitable method	3 days	9/8/17 8:00 AM	9/12/17 5:00 PM	5
7		Extract Indian tweets	10 days	9/13/17 8:00 AM	9/26/17 5:00 PM	6
8		Preprocess the data	29 days	9/27/17 8:00 AM	11/6/17 5:00 PM	7
9		Tokenization	5 days	9/27/17 8:00 AM	10/3/17 5:00 PM	7
10		Remove Twitter Notations	4 days	10/4/17 8:00 AM	10/9/17 5:00 PM	9
11		Remove URLs,hyperlinks,emoticons	4 days	10/10/17 8:00 AM	10/13/17 5:00 PM	10
12		Remove stop words	4 days	10/16/17 8:00 AM	10/19/17 5:00 PM	11
13		Compress elongated words	6 days	10/20/17 8:00 AM	10/27/17 5:00 PM	12
14		Decompress the slang words	6 days	10/30/17 8:00 AM	11/6/17 5:00 PM	13
15		Extract features	6 days	11/7/17 8:00 AM	11/14/17 5:00 PM	14
16		Identify suitable feature extraction method	2 days	11/7/17 8:00 AM	11/8/17 5:00 PM	14
17		Extract corresponding aspects	4 days	11/9/17 8:00 AM	11/14/17 5:00 PM	16
18		Classify based on sentiment	30 days	11/15/17 8:00 AM	12/26/17 5:00 PM	17
19		Collect training data	3 days	11/15/17 8:00 AM	11/17/17 5:00 PM	17
20		Evaluate classification techniques	3 days	11/20/17 8:00 AM	11/22/17 5:00 PM	19
21		Select suitable technique	2 days	11/23/17 8:00 AM	11/24/17 5:00 PM	20
22		Build classifier	10 days	11/27/17 8:00 AM	12/8/17 5:00 PM	21
23		Train classifier using training data	5 days	12/11/17 8:00 AM	12/15/17 5:00 PM	22
24		Classify test data using trained model	7 days	12/18/17 8:00 AM	12/26/17 5:00 PM	23
25		Evaluate results	7 days	12/27/17 8:00 AM	1/4/18 5:00 PM	24
26		Measure correlation of results with official food price inflation da..	5 days	1/5/18 8:00 AM	1/11/18 5:00 PM	25
27		Monitor real-time food related economic indicators	30 days	1/12/18 8:00 AM	2/22/18 5:00 PM	26

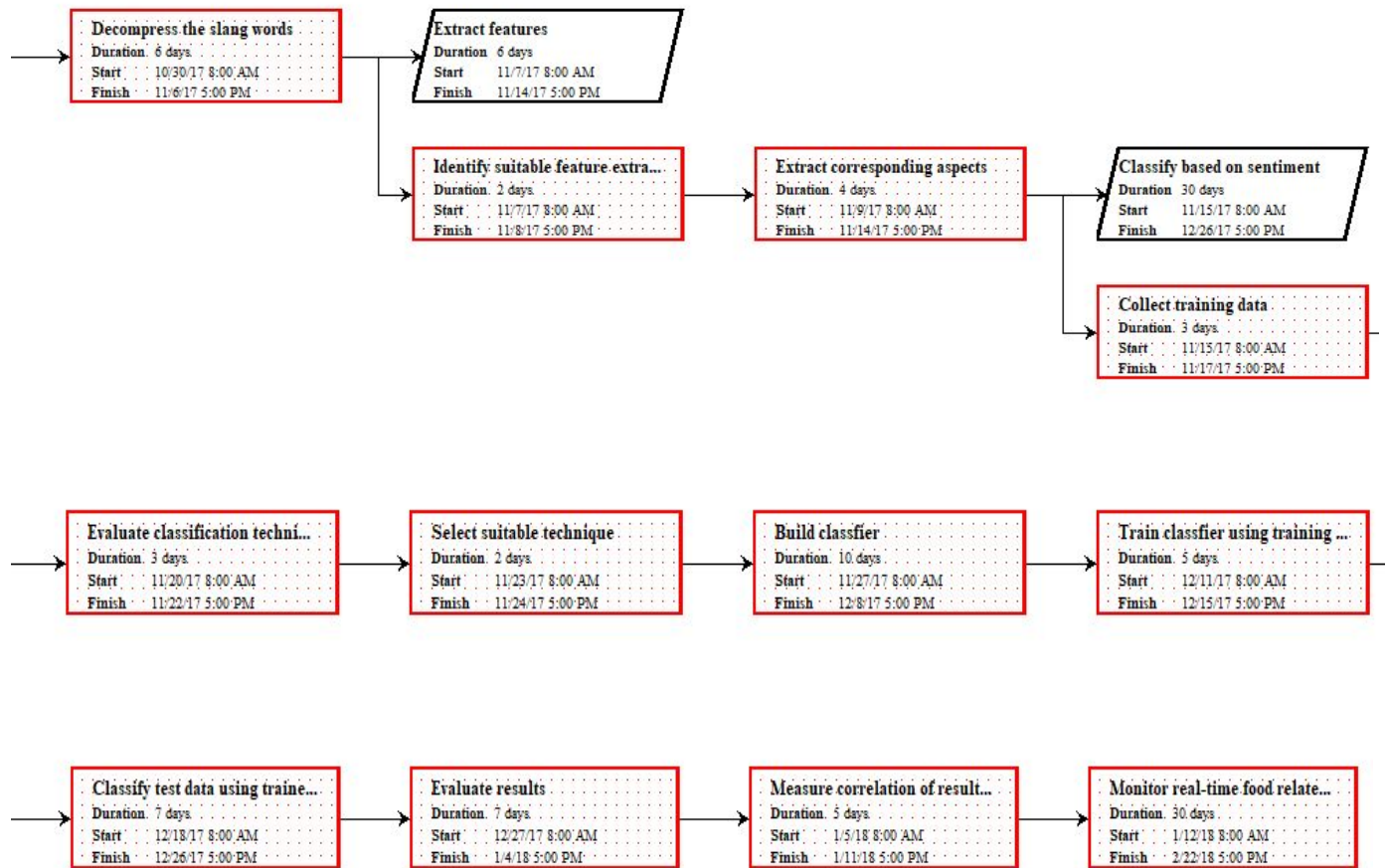


			11 Feb 18					18 Feb 18					25 Feb 18					4 Mar 18					11 Mar 18					18 Mar 18									
T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S

4. Pert chart

PERT stands for program evaluation review technique. Pert chart is a project management tool which is used to organise, schedule and also coordinate the tasks in a project. It is a graphic illustration of the project. The following chart has been generated using Projectlibre to schedule the tasks of our project. The first three tasks are sequential followed by two concurrent tasks. As the name suggests sequential tasks must be completed in the same order or sequence. Concurrent tasks are not dependent on the completion of one to start the other.





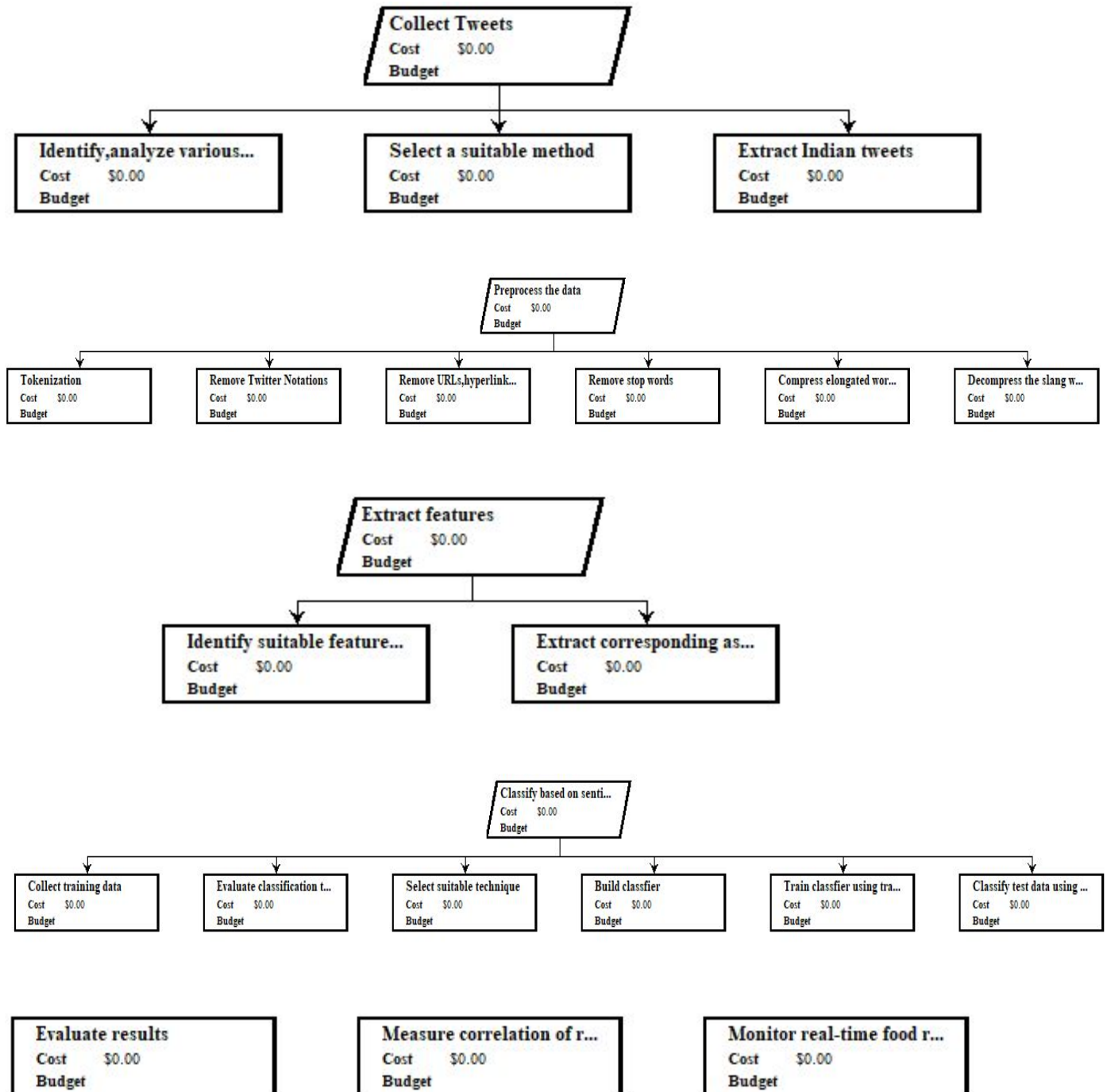
5. WBS

WBS stands for work breakdown structure. It is a key project deliverable that conveys the hierarchical decomposition of work which is to be executed by the project team. It organises team's work into manageable sections.

Identify Team
Cost \$0.00
Budget

Identify Problem State...
Cost \$0.00
Budget

Identify keywords
Cost \$0.00
Budget



6. CPM Chart

CPM stands for critical path method. It is a step by step project management technique that defines critical and noncritical tasks. It's goal is to prevent timeframe problems and process bottlenecks. In this method, critical activities will have a direct impact on the completion date of the project. This chart helped us in identifying the critical tasks of our project.

Task Information

ID	Name	Start	Finish	Critical	Late Start	Late Finish	Free Slack	Total Slack
1	Identify Team	8/16/17 8:00 AM	8/16/17 5:00 PM	true	8/16/17 8:00 AM	8/16/17 5:00 PM	0 days	0 days
2	Identify Problem Statement	8/17/17 8:00 AM	8/25/17 5:00 PM	true	8/17/17 8:00 AM	8/25/17 5:00 PM	0 days	0 days
3	Identify keywords	8/28/17 8:00 AM	8/30/17 5:00 PM	true	8/28/17 8:00 AM	8/30/17 5:00 PM	0 days	0 days
4	Collect Tweets	8/31/17 8:00 AM	9/26/17 5:00 PM	true	8/31/17 8:00 AM	9/26/17 5:00 PM	0 days	0 days
5	Identify,analyze various tweet	8/31/17 8:00 AM	9/7/17 5:00 PM	true	8/31/17 8:00 AM	9/7/17 5:00 PM	0 days	0 days
6	Select a suitable method	9/8/17 8:00 AM	9/12/17 5:00 PM	true	9/8/17 8:00 AM	9/12/17 5:00 PM	0 days	0 days
7	Extract Indian tweets	9/13/17 8:00 AM	9/26/17 5:00 PM	true	9/13/17 8:00 AM	9/26/17 5:00 PM	0 days	0 days
8	Preprocess the data	9/27/17 8:00 AM	11/6/17 5:00 PM	true	9/27/17 8:00 AM	11/6/17 5:00 PM	0 days	0 days
9	Tokenization	9/27/17 8:00 AM	10/3/17 5:00 PM	true	9/27/17 8:00 AM	10/3/17 5:00 PM	0 days	0 days
10	Remove Twitter Notations	10/4/17 8:00 AM	10/9/17 5:00 PM	true	10/4/17 8:00 AM	10/9/17 5:00 PM	0 days	0 days
11	Remove URLs,hyperlinks,	10/10/17 8:00 AM	10/13/17 5:00 PM	true	10/10/17 8:00 AM	10/13/17 5:00 PM	0 days	0 days
12	Remove stop words	10/16/17 8:00 AM	10/19/17 5:00 PM	true	10/16/17 8:00 AM	10/19/17 5:00 PM	0 days	0 days
13	Compress elongated words	10/20/17 8:00 AM	10/27/17 5:00 PM	true	10/20/17 8:00 AM	10/27/17 5:00 PM	0 days	0 days
14	Decompress the slang words	10/30/17 8:00 AM	11/6/17 5:00 PM	true	10/30/17 8:00 AM	11/6/17 5:00 PM	0 days	0 days
15	Extract features	11/7/17 8:00 AM	11/14/17 5:00 PM	true	11/7/17 8:00 AM	11/14/17 5:00 PM	0 days	0 days
16	Identify suitable feature extraction	11/7/17 8:00 AM	11/8/17 5:00 PM	true	11/7/17 8:00 AM	11/8/17 5:00 PM	0 days	0 days
17	Extract corresponding aspects	11/9/17 8:00 AM	11/14/17 5:00 PM	true	11/9/17 8:00 AM	11/14/17 5:00 PM	0 days	0 days
18	Classify based on sentiment	11/15/17 8:00 AM	12/26/17 5:00 PM	true	11/15/17 8:00 AM	12/26/17 5:00 PM	0 days	0 days
19	Collect training data	11/15/17 8:00 AM	11/17/17 5:00 PM	true	11/15/17 8:00 AM	11/17/17 5:00 PM	0 days	0 days
20	Evaluate classification techniques	11/20/17 8:00 AM	11/22/17 5:00 PM	true	11/20/17 8:00 AM	11/22/17 5:00 PM	0 days	0 days
21	Select suitable technique	11/23/17 8:00 AM	11/24/17 5:00 PM	true	11/23/17 8:00 AM	11/24/17 5:00 PM	0 days	0 days
22	Build classifier	11/27/17 8:00 AM	12/8/17 5:00 PM	true	11/27/17 8:00 AM	12/8/17 5:00 PM	0 days	0 days
23	Train classifier using training data	12/11/17 8:00 AM	12/15/17 5:00 PM	true	12/11/17 8:00 AM	12/15/17 5:00 PM	0 days	0 days
24	Classify test data using trained	12/18/17 8:00 AM	12/26/17 5:00 PM	true	12/18/17 8:00 AM	12/26/17 5:00 PM	0 days	0 days
25	Evaluate results	12/27/17 8:00 AM	1/4/18 5:00 PM	true	12/27/17 8:00 AM	1/4/18 5:00 PM	0 days	0 days
26	Measure correlation of results	1/5/18 8:00 AM	1/11/18 5:00 PM	true	1/5/18 8:00 AM	1/11/18 5:00 PM	0 days	0 days
27	Monitor real-time food related	1/12/18 8:00 AM	2/22/18 5:00 PM	true	1/12/18 8:00 AM	2/22/18 5:00 PM	0 days	0 days

7. Team skills

S.no	Name	Work done	Work to be done
1	Sandhya	Identifying problem statement and keywords Identifying various tweet collection methods Tokenization of tweets Removal of hyperlinks,URL's and emoticons from tokens Removal of stopwords from tokens	Extract corresponding aspects Collect training data Evaluation and selection of classification techniques Building and training classifier Classification of test data using trained model Evaluation of results Monitor real time food related economic indicators
2	Tejaswi	Identifying problem statement and keywords Extraction of tweets Tokenization of tweets Removal of twitter notations Compression of elongated words	Identify suitable feature extraction method Collect training data Evaluation and selection of classification techniques Building and training classifier Classification of test data using trained model Evaluation of results Monitor real time food related economic indicators
3	Saranya	Identifying problem statement and keywords Extraction of corresponding aspects Collect training data Evaluation and selection of classification techniques	Decompression of slang words Identify suitable feature extraction method Collect training data Evaluation and selection of classification techniques Building and training classifier Classification of test data using trained model

			Evaluation of results Monitor real time food related economic indicators
--	--	--	---

8. Conclusion

Social media has become an incredible source of public opinion and sentiments. Mining this data and monitoring it can help the government to get a faster and better idea of the statistics than traditional methods which will in turn result in taking necessary measures faster. Work has been divided equally among team members based on their capabilities and interests in each task. ProjectLibre has been useful in planning and organizing tasks in this project in an efficient way thereby reducing risks. Charts generated using the tool help in planning the future tasks accordingly so that deadlines are met.