**CERTIFICATE**

This is to certify that this project work entitled **Text-Mining Twitter Data Using Python and R,** submitted in partial fulfillment to the requirements for the award of the Master of Science in Computer Science and Statistics is a record of bonafied work carried out by the following students during the year 2014-2015.

G. Siva Kumar (Regd No.:713212038005)

S. Vyshamya (Regd No.:713212038008)

P. Kavitha (Regd No.:713212038009)

K. Papa Laxmi (Regd No.:713212038019)

D. Sravani Kumari (Regd No.:713212038020)

Dr. L. VENKATESWARA RAO Dr. B. MUNYSWAMY

Associate Professor, Head of the Department,

Dept. of Statistics, Associate Professor,

Andhra University, Dept. of Statistics,

Visakhapatnam – 530 003 Andhra University,

Visakhapatnam – 530 003

**DECLARATION**

We hereby declare that this project work entitled **Text-Mining Twitter Data Using Python and R**, is authentic work carried out for the partial fulfillment to the requirements for the award of the Master of Science in Computer Science and Statistics Degree and we further admit that this has not been submitted by us anywhere else for the award of any other degree.

G. Siva Kumar S. Vyshamya

(Regd No.:713212038005) (Regd No.:713212038008)

P. Kavitha K. Papa Laxmi

(Regd No.:713212038009) (Regd No.:713212038019)

D. Sravani Kumari

(Regd No.:713212038020)

**ACKNOWKEDGEMENT**

We express our sincere and special thanks to our teacher and project guide Dr. L. Venkateswra Rao, Department of Statistics, Andhra University for his esteemed guidance, valuable suggestions and encouragement in completing this project successfully.

We thank – Dr. B. Muniswamy, Head, Department of Statistics, Andhra University for his constant support.

We also thank – Prof. K. Srinivasa Rao, Chairman, P.G. Board of Studies, Andhra University, who is always a source of inspiration.

We express our special thanks to our friends Mr. Siva Krishna and Mr. Venkatesh for sparing their laptops whenever we wanted to demonstrate our progress of work to our project guide. We also thank all our friends for their suggestions and excellent help during the course of this project.

We are all grateful to our parents and family members for their continuous support and encouragement throughout the completion of our project.

**Index**

**Topics**  **Page.No**

**Chapter 1: Introduction to Sentiment Analysis 1-11**

* 1. **Need of Sentiment Analysis. 02**
  2. **New Approaches to carryout sentiment analysis. 03**
  3. **Applications of sentiment analysis 03**
  4. **How to carry out sentiment analysis 04**
  5. **Sentiment Classification Techniques 05**
     1. Machine learning approach 06
     2. Lexicon-Based Method of Classification 07
  6. **Objectives 08**
  7. **Sources of data 08**
     1. **Twitter 08**
     2. History and Growth of Twitter 09
     3. Some Features of twitter 09
     4. **Face book 11**
  8. **Description of Chapters 11**

### **Chapter 2: Application Programming tools 12-20**

### **2.1 Application Programming Interface 13**

### 2.1.1 Twitter Application programming Interface 13

### 2.1.2 How to create Twitter an Application 14

### 2.1.3 Twitter API Access 16

2.1.4 Twitter API limits 16

**2.2 Programming tools 16**

2.2.1 Introduction to Python 16

2.2.2 Introduction to R 19

2.2.3 Packages in R 20

**Chapter 3: Sentiment analysis using Twitter data 21-42**

**3.1 Data collection using Twitter Streaming API and Python 23**

**3.2 Preprocessing of a Twitter Data 25**

3.2.1Steps for data cleaning 25

* 1. **Word consideration 31**

3.3.1 Unigram 31

* + 1. Bigram 32

**3.4 Scoring module 32**

3.4.1 AFINN lexicon 32

3.4.2 NRC Hashtag Sentiment Lexicon 32

3.4.3 Scoring method 33

**3.5 Data Visualization and Interpretation 36**

3.5.1 Wordcloud 37

3.5.2 Multiple Bar graphs 40

**Chapter 4: Cluster Analysis in Text Mining using R-packages 43-58**

**4.1 Introduction to cluster Analysis 44**

**4.2 Preprocessing using R 45**

4.2.1Retrieving Text from Twitter 46

4.2.2 Transforming Text 46

4.2.3 Creating Plain Text Document words 47

4.2.4 Building a Term-Document matrix 48

4.2.5 Frequent terms and Associations 49

**4.3 Clustering of Words Using Hierarchical Clustering 52**

4.3.1 Clustering of iphone words 53

4.3.2 Clustering of Motorola words 55

4.3.3 Clustering of Nokia words 56

4.3.4 Clustering of Samsung words 57

**Chapter 5: Clustering Of Tweets Using Partitioning Methods 59-84**

**5.1 Clustering of Tweets Using K-Means Algorithm 60**

5.1.1 Applying k-means algorithm using R language 61

5.1.2 Clustering of iphone tweets 62

5.1.3 Clustering of Motorola tweets 65

5.1.4 Clustering of Nokia tweets 68

5.1.5 Clustering of Samsung tweets 71

**5.2 Clustering of Tweets Using K-Medoids Algorithm 74**

5.2.1 Applying k-Medoids algorithm using R language 74

5.2.2 Clustering of iphone tweets 75

5.2.3 Clustering of Motorola tweets 77

5.2.4 Clustering of Nokia tweets 80

5.2.5 Clustering of Samsung tweets 82

**Bibliography 85**