**M S Ramaiah Institute of Technology**

(An Autonomous Institute, Affiliated to VTU)

MSR nagar, MSRIT post, Bangalore-54

A Dissertation Report on

**Data Analysis on Twitter Data**

Submitted by

Praveen K 1MS12CS075

Praveengouda S Patil 1MS12CS076

Manoj A 1MS12CS053

Jayachitra T R 1MS12CS040

*in partial fulfillment for the award of the degree of*

# *Bachelor of Engineering in Computer Science & Engineering*



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**M.S.RAMAIAH INSTITUTE OF TECHNOLOGY**

**(Autonomous Institute, Affiliated to VTU)**

**BANGALORE-560054**

[www.msrit.edu](http://www.msrit.edu), **May 2015**

**Acknowledgements**

First and foremost,my utmost Gratitude to our HOD, Dr. K G Srinivasa who gave us a chance to complete this project. He has been our inspiration as we overcome all the obstacles in the completion of this project work. He had kind concern and consideration regarding project work which helped us to successfully complete the project.

Secondly we would like to thank Sowmya B J, Assistant Professor, MSRIT and Chandrika Prasad, Assistant Professor who reviewed our project and gave us a valuable guidance which helped us in completing this project.

This work would not have been possible without the guidance and help of several individuals who in one way or another contributed their valuable assistance in preparation and completion of this study.

We would like to express sincere thanks to all the teaching and non-teaching faculty of CSE department and my dear friends who helped in all the ways while preparing the Report.

# Abstract

Twitter, one of the largest social media site receives tweets in millions every day. This huge amount of raw data can be used for industrial or business purpose by organizing according to our requirement and processing. This paper provides a way of analysing tweeter data using hadoop which will process the huge amount of data on a hadoop cluster faster in real time.

The main purpose of this project is to collect large data from social networking site 'TWITTER' and analyse this data which can depict the public sentiments and emotions that tells the nature of public .The aim of this project is to come up with a good result that helps the public from the analysis of user tweets. This includes identifying the locations where the tweets are made more and differentiating between positive, negative and neutral tweets.

#### Contents

***Declaration i***

***Acknowledgements ii***

***Abstract iii***

1 **INTRODUCTION**

* 1. General Introduction
  2. Statement of the Problem
  3. Objectives of the project
  4. Project deliverables
  5. Current Scope
  6. Future Scope

1. **PROJECT ORGANIZATION**
   1. Software Process Models
   2. Roles and Responsibilities
2. **LITERATURE SURVEY**

3.1 Introduction

3.2 Main Body

3.3 Conclusion of Survey

1. **SOFTWARE REQUIREMENT SPECIFICATIONS**

4.1 Product Overview

4.2 External Interface Requirements

4.2.1 User Interfaces

4.2.2 Hardware Interfaces

4.2.3 Software Interfaces

4.2.4 Communication Interfaces

4.3 Functional Requirements

4.3.1 Functional Requirement 1.1

:

4.3.n Functional Requirement 1.n

4.4 Software System Attributes

4.4.1 Reliability

4.4.2 Availability

4.4.3 Security

4.4.4 Portability

4.4.5 Maintainability

4.5 Performance Requirements

4.6 Database Requirement

4.7 Design Constraints

4.8 Other Requirements

1. **DESIGN**
   1. Introduction
   2. Architecture Design
   3. Graphical User Interface
   4. Class Diagram
   5. Sequence Diagram
   6. Data flow diagram
   7. Metric calculation
2. **IMPLEMENTATION**
   1. Tools used
   2. Technology used
   3. Overall view of the project in terms of implementation
   4. Explanation of Algorithm and how it is been implemented
   5. Information about the implementation of Modules
3. **TESTING**

**7.1** Results and Snapshots

1. **CONCLUSION & SCOPE FOR FUTURE WORK**
2. **REFERENCES**
3. **INTRODUCTION**
   1. General Introduction

The rise of micro blogging services like Twitter has spawned great interest in these systems as human-powered sensing networks. Since its creation in 2006, Twitter has experienced an exponential explosion in its user base, reaching a lot of people across the globe. So analysis on the twitter data will help the people in knowing the mindset of the people around us.

Hence in this project we are doing an overall analysis of the twitter dataset which will classify the tweets as positive, negative or neutral. The main technologies used in this are eclipse, Hadoop and map reduce parallel programming concept. The detailed design and implementation is explained below. Here we concentrate on the analysing the dataset and visualising the output using pie chart which is done using python.

* 1. Problem Statement

The purpose of this project is to find out trends by aggregating the data in social networking site such as Twitter. Analysis of Twitter data to evaluate sentiment of the user tweets and classify them as positive, negative or neutral tweets.

1.3 Objectives of the Project

The objective is to implement an algorithm for automatic classification of text into positive, negative or neutral. Sentiment Analysis to determine the attitude of the mass is positive, negative or neutral towards the subject of interest and Graphical representation of the sentiment in form of Pie-Chart.

1.4 Project Deliverables

The PPT has been prepared and the document has been made on this topic and same thing has been presented which explains the above mentioned things.

The SRS has been prepared and the Design document has been made on this topic and same thing has been presented which explains the above mentioned things.

The output is being visualized using Pie-chart by classifying the tweets as positive, negative and neutral.

1.5 Current Scope

Twitter has lot of scope in the modern era. Twitter as a social media has many users and the number of users is increasing day by day. At this moment, the code can handle the analysis part with a very good accuracy. But there are a few areas which have a lot of scope in this aspect. Sarcastic comments are the ones which are very difficult to identify. Tweets containing sarcastic comments give exactly opposite results owing to the mindset of the author. These are almost impossible to track. Also depending on the context in which a word is used, the interpretation changes. For ex: the word ‘unpredictable’ in ‘unpredictable plot’ in context of a land plot is negative whereas ‘unpredictable plot ’ in context of a movie’s plot is positive. So it’s important to relate the interpretation with the context of the tweets. Also the use of native language combined with English usage is difficult to interpret.

1.6 Future Scope

Nowadays big data has become the buzzword in IT industry organizations. The need of analysing and processing of information has grown a lot. This paper implemented the analysing of big data (tweets) only for text. Further analysis can be done to images and all types of multimedia files based on index support. The result of Text mining and data analysis would help in suggesting related pages based on different types of data. So that industries make the data easily available to people who is using and trying accessing such type of data.

1. **PROJECT ORGANIZATION**

2.1 Software Process Models

**Agile Model**: We choose to follow it based on the project size, complexity and duration. The agile methodology is a defined process that makes it easier for a project team to deal with complexity and coherence using an incremental delivery approach. If project requirements are clearly defined and understood and expected project duration is 3 to 4 months and hence we suggest iterative and incremental development model with periodic reviews.



This model reduces the project development risk because of multiple incremental elaboration, construction cycle and addition of new features. In addition to four basic phases such as inception, elaboration, construction and transition, we provide support for project specific timeframe.

2.2 Roles and Responsibilities

The roles and responsibilities for the project are-

* Praveen K- Maintaining the backend (Hadoop) and linking the functionality to the frontend. Technologies used are Hadoop, Mapreduce and Java.
* Manoj A- Normalizing the dataset, stemming and implementing Naïve Bayes algorithm. Technologies used are Porter’s algorithm, Naïve Bayes algorithm and Java.
* Jayachitra T R- Collecting the dataset, performing the tests, documenting the details. Tehnologies used are Java, Hadoop and Mapreduce.
* Praveen S P- Classifying the dataset as positive, negative and neutral using python, installation of hadoop and monitoring the project. Technologies used are Python, Hadoop and Java.

1. **LITERATURE SURVEY**

3.1 Introduction

The rise of micro blogging services like Twitter has spawned great interest in these systems as human-powered sensing networks. Since its creation in 2006, Twitter has experienced an exponential explosion in its user base, reaching a lot of people across the globe. So analysis on the twitter data will help the people in knowing the mindset of the people around us.

Hence in this project we are doing an overall analysis of the twitter dataset which will classify the tweets as positive, negative or neutral. The main technologies used in this are eclipse, Hadoop and map reduce parallel programming concept. The detailed design and implementation is explained below. Here we concentrate on the analysing the dataset and visualising the output using pie chart which is done using python.

3.2 Main Body

Hadoop supports the MapReduce model, which was introduced by Google as a method of solving a class of petascale problems with large clusters of inexpensive machines. The model is based on two distinct steps for an application:

• Map:

An initial ingestion and transformation step, in which individual input records can be processed in parallel.

• Reduce:

An aggregation or summarization step, in which all associated records, must be processed together by a single entity.

The core concept of MapReduce in Hadoop is that input may be split into logical chunks, and each chunk may be initially processed independently, by a map task. The results of these individual processing chunks can be physically partitioned into distinct sets, which are then sorted. Each sorted chunk is passed to a reduce task. So In this project we have this concept to classify the nature of tweets.

3.3 Conclusion of Survey

This project gave us hands on experience of handling and parallel processing of huge amount of data. Data collection process introduced us to java twitter streaming API. It was very interesting to gather and then aggregate the social networking data so as to extract interesting patterns and recent trends from it. We got exposure to work with prominent parallel data processing tool: Hadoop, Apache Hadoop framework is gaining significant momentum from both industry and academia as the volume of data to analyze growth rapidly.

1. **SOFTWARE REQUIREMENT SPECIFICATIONS**

4.1 Product Overview

The project involves building of a system that can be used to predict the nature of tweets. It is implemented using Hadoop as a platform where the dataset is stored and uses Mapreduce parallel programming concept.

The final output is visualized using a Pie-chart where ot shows the classification of tweets as positive, negative and neutral. This is implemented using Python. This helps in understanding the nature of tweets and the nature of the people in our society.

* 1. External Interface Requirements

4.2.1 User Interfaces:

The interface will meet the following requirements to conform to the users’ needs. It will be simple and easy to understand. Controls which allow the user to interact with the application will be clear and imply their functionality within the application.

The user interface of eclipse was simple to navigate and easy to understand.

* + 1. Hardware Interfaces:

The application will run on a password protected personal Microsoft laptop. No further hardware devices or interfaces will be required for this analysis.

* + 1. Software Interfaces:

**Inputs**

The java program takes the dataset, positive and negative words as the input.

**Outputs**

Graphical representation of the positive and negative tweets corresponding to various locations.

4.2.4 Communication Interfaces:

Internet connection and a web browser are required in order to make use of several functions and to be executed such as searching, viewing and downloading.

4.3 Functional Requirements

4.3.1 Retrieving Input:

The software will receive three inputs:

* Userid.
* TweetId.
* Tweets.

4.3.2 Data set pre-Processing:

Data set is pre-processed using eclipse IDE to eliminate stop word and stemming is done using porter’s algorithm.

4.3.3 Tweet classification:

Tweet classification will be performed on the user Tweet to determine the nature of the Tweet relative to the geolocation. The data analysis will provide a negative, neutral, or positive numeric value.

* + 1. Output :

The output is a real time data in the form of a simple graph. The graph will illustrate the nature of tweets over different geolocations. This output should be clear and easy to understand.

4.4 Software System Attributes

4.4.1 Reliability:

The project will meet all of the functional requirements without any unexpected behavior. At no time will the graph display incorrect information.

4.4.2 Availability:

The project will be available on demand. The functionality of the project will not depend on any external services such as internet access that are required.

4.4.3 Security:

The project will never disclose any personal information of Twitter users, and will collect no personal information from its own users.

4.4.4 Portability:

This project is designed to run on any operating system.

4.4.5 Maintainability:

The project should be written clearly and concisely. The code will be well documented. Particular care will be taken to design the project modularly to ensure that maintenance is easy.

4.5 Performance Requirements

4.5.1 Real-Time:

The project will provide up-to-date information, limited only by the rate of Twitter input. The graph output will display the latest results at all times.

4.5.2 System Resource Consumption:

Resource consumption of this application should not reach an amount that affects the normal process of system. The application should be capable of operating in the background.

2GB of Ram is required and the system storage of more than 8GB is required.

4.6 Database Requirement

The Hadoop S3 database is used to store the huge datasets.

4.7 Design Constraints

The constraints that affect the design of this project are the cost and the performance. The design of this project is cost-effective and the performance is high as parallel programming is used using HPC tool like Hadoop.

4.8 Other Requirements

Knowledge on Twitter social media, tweet’s attributes, Internet Access to World Wide Web.

1. **DESIGN**

5.1 Introduction

5.1.1 Number of Modules: 4

5.1.2 Modules Description:

* **Fetching and Extracting Data:**

The data set contains training set and test set. The training set contains 115,886 twitter users and 1,19,629 tweets from the users. The test set contains 5,136 twitter users and 1,22,891 tweets from the users. All the locations of users are uploaded from their smart phones with the form of "UT: Latitude , Longitude".

* **Data Preprocessing:**

Data set is pre-processed using eclipse IDE to eliminate stop word and stemming is done using porter’s algorithm.

* **Classification:**

Tweet classification will be performed on the user Tweet to determine the nature of the Tweet relative to the geolocation. The data analysis will provide a negative, neutral, or positive numeric value.

* **Analysis:**

The Preprocessed data is considered for analysis. Based on the number of positive and negative words in the user tweets, the tweet is classified as positive, negative or neutral tweets and the same is represented graphically.

5.1.3Algorithm Design:

* **Porter’s Algorithm (stemming)**

Step 1: Gets rid of plurals and -ed or -ing suffixes.

Step 2: Turns terminal y to i when there is another vowel in the stem.

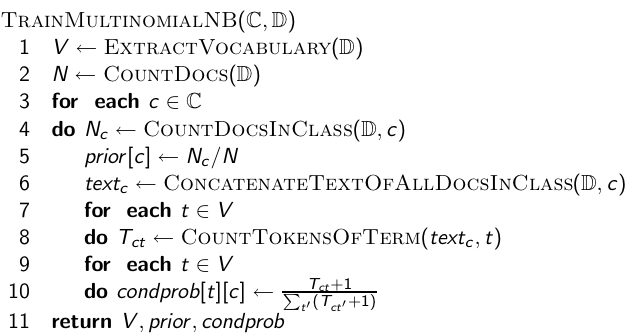
Step 3: Maps double suffixes to single ones: -ization, -ational, etc.

Step 4: Deals with suffixes, -full, -ness etc.

Step 5: Takes off -ant, -ence, etc.

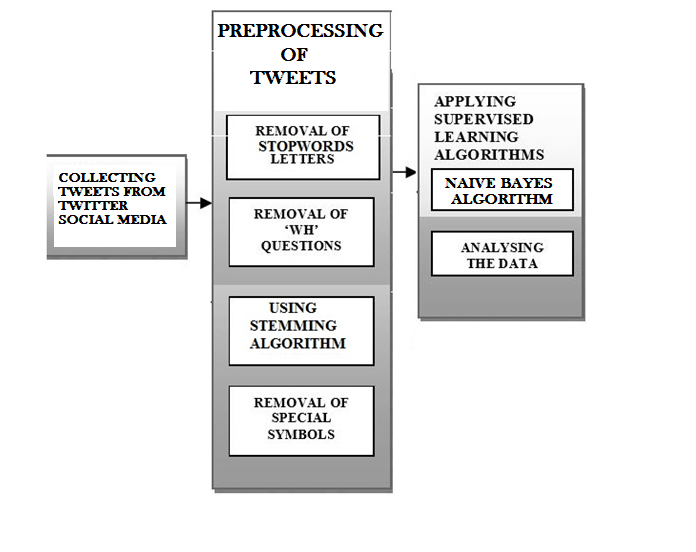
Step 6: Removes a final –e.

* **Naive bayes algorithm (classification)**

******

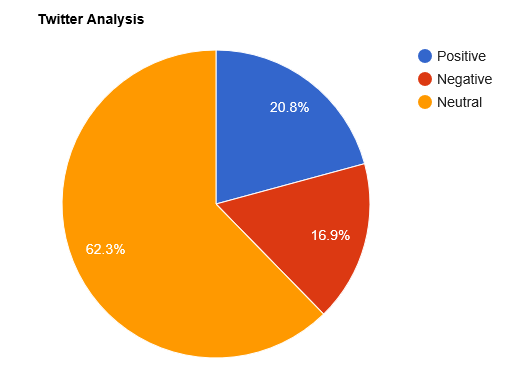
**Example: **

5.2 Architecture Design



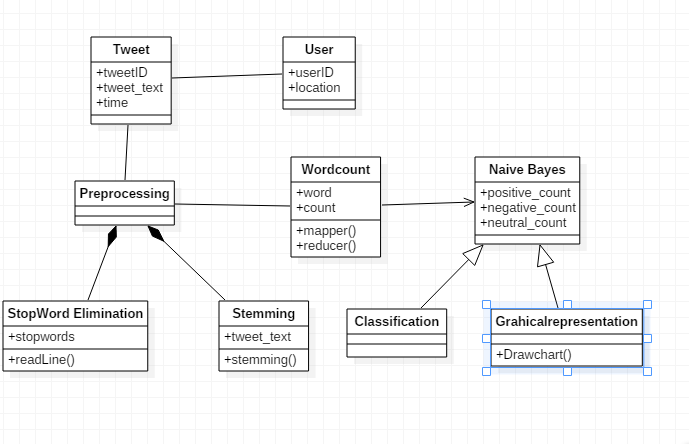
5.3 Graphical User Interface

The interface will meet the following requirements to conform to the users’ needs. It will be simple and easy to understand. Controls which allow the user to interact with the application will be clear and imply their functionality within the application. The user interface of eclipse was simple to navigate and easy to understand.

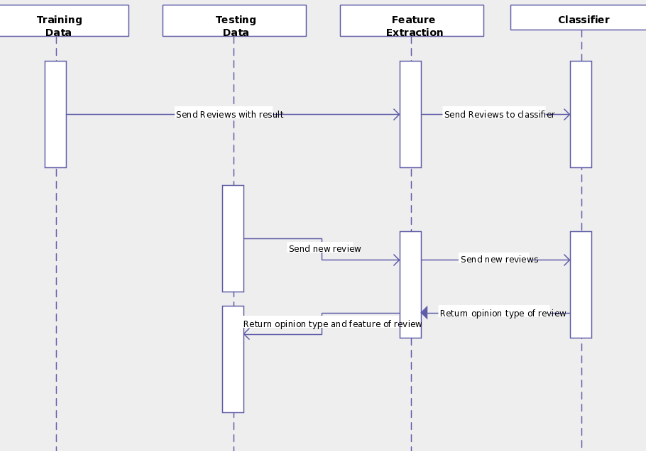


The output is a real time data in the form of a simple graph. The graph will illustrate the nature of tweets over different geolocations. This output should be clear and easy to understand.

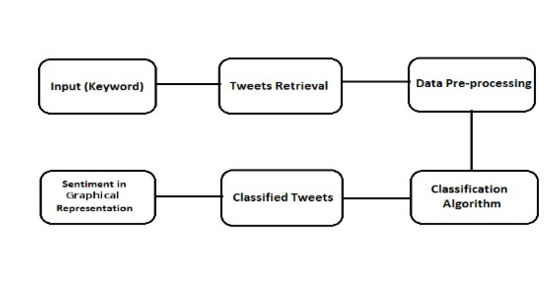
5.4 Class Diagram



5.5 Sequence Diagram



5.6 Data flow diagram



**6. IMPLEMENTATION**

6.1 Tools Used:

Hadoop, Eclipse, Java, Idle

6.2 Technology Used:

Mapreduce concept, Python, Java

6.3 Overall view of the project in terms of implementation

The Project is implemented as follows:

* Twitter Data Set Identification and Extraction
* Pre-processing of data set
  + Removing unwanted attributes
  + Eliminating stop words
  + Stemming ( Porter’s algorithm)
* Tweet classification
  + Positive
  + Negative
  + Neutral

6.4 Explanation of Algorithm and how it is been implemented

**Our Approach In Implementation**:

In our approach we focused more on the speed of performing analysis than its accuracy i.e. performing sentiment analysis on big data which is achieved by splitting the various modules of data in following steps and collaborating with hadoop for mapping it onto different machines. Part of speech tagged using opennlp. This tagging is used for following various purposes.

i. **Stop words removal:**

The stop words like a, an, this which are not useful in performing the analysis are removed in this phase. Stop words are removed using java in eclipse platform. All the words are not considered are not considered for analysis.

ii. **Unstructured to structured**:

Twitter comments are mostly unstructured i.e. ‘aswm’ is written ‘awesome’, ‘happyyyyyy’ to actually ‘happy’. Conversion to structured is done by dynamic data records of unstructured to structured and vowels adding.

iii. **Stemming**:

In stemming we use porter’s algorithm to remove suffix and prefixes from the tweeter data set file and normalise it.

iv. **Map reduce using Hadoop:**

This is used for classifying the test\_set\_tweets file into positive negative and neutral by using parallel processing in hadoop using map reduce.

6.5 Information about the implementation of Modules

Number of Modules: 4

* Fetching and Extracting Data:
* Data Pre-processing:
* Classification:
* Analysis

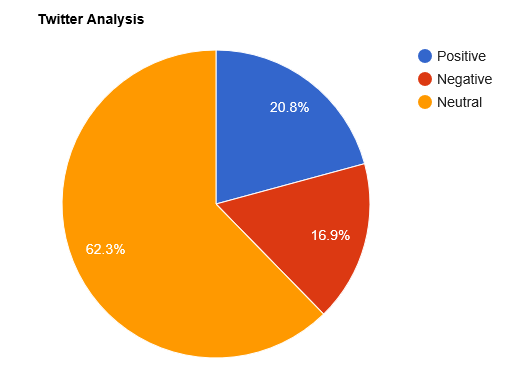
Algorithm Design:

Porter’s Algorithm

**7. TESTING**

7.1 Results and Snapshots

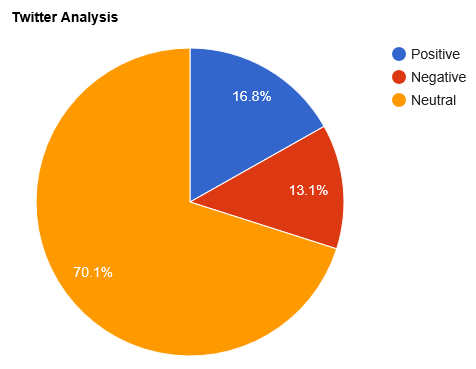
After streaming the tweets into HDFS in real time, Naïve bayes is used in analyzing the tweets. Tweets are tagged as documents where categories are the hash tags defined in the configuration file. Later the tweets are grouped as positive, negative and neutral based on subjectivity corpus forming a dictionary of words and its polarity. The sample example shown in pie chart which is obtained by running the map reduce code in hadoop.

**GRAPHICAL REPRESENTATION OF NATURE OF USER** **TWEETS**  
 

The sample count or the given above chart is shown in the following table:

|  |  |
| --- | --- |
| Opinion | Count |
| Positive | 25540 |
| Negative | 20765 |
| Neutral | 76586 |

**GRAPHICAL REPRESENTATION OF NATURE OF SINGLE USER** **TWEET**



The sample count or the given above chart is shown in the following table:

|  |  |
| --- | --- |
| Opinion | Count |
| Positive | 313 |
| Negative | 152 |
| Neutral | 555 |

**8. CONCLUSION & SCOPE FOR**

**FUTURE WORK**

This project gave us hands on experience of handling and parallel processing of huge amount of data. Data collection process introduced us to java twitter streaming API. It was very interesting to gather and then aggregate the social networking data so as to extract interesting patterns and recent trends from it. We got exposure to work with prominent parallel data processing tool: Hadoop. Apache Hadoop framework is gaining significant momentum from both industry and academia as the volume of data to analyze growth rapidly.

This project helped us not only to gain knowledge about installation and configuration of hadoop distributed file system but also map reduce programming model. At the end of analysis phase data visualization was performed with the help of Google Developer. Amongst the many fields of analysis, there is one field where humans have dominated the machines more than any – the ability to analyze sentiment, or sentiment analysis.

Twitter has lot of scope in the modern era. Twitter as a social media has many users and the number of users is increasing day by day. At this moment, the code can handle the analysis part with a very good accuracy. But there are a few areas which have a lot of scope in this aspect. Sarcastic comments are the ones which are very difficult to identify. Tweets containing sarcastic comments give exactly opposite results owing to the mindset of the author. These are almost impossible to track. Also depending on the context in which a word is used, the interpretation changes. For ex: the word ‘unpredictable’ in ‘unpredictable plot’ in context of a land plot is negative whereas ‘unpredictable plot’ in context of a movie’s plot is positive. So it’s important to relate the interpretation with the context of the tweets. Also the use of native language combined with English usage is difficult to interpret.

**9. REFERENCES**

* <https://archive.org/download/twitter_cikm_2010>
* <http://www.cloudera.com/content/www/enus/documentation/other/tutorial/CDH5/Hadoop-Tutorial/ht_example_4_sentiment_analysis.html>
* <https://github.com/omarshammas/sentiment_analysis>
* <https://github.com/timvandermeij/sentiment-analysis/tree/master/words>
* <https://github.com/madhusudancs/sentiment-analyzer/blob/master/analyzer/train.py>
* <http://academictorrents.com/details/d8b3a315172c8d804528762f37fa67db14577cdb>
* <https://archive.org/details/twitter_cikm_2010>
* <http://www.mrgeek.me/technology/datascience/data-mining-1-5-million-tweets-for-twitter-sentiment-analysis/>
* <http://www.slideshare.net/sumit786raj/sentiment-analysis-of-twitter-data?related=1>
* <http://www.slideshare.net/niteshsinghns/twitter-sentiment-analysis-project-report>
* <http://www.alex-hanna.com/tworkshops/lesson-6-basic-sentiment-analysis/>
* <http://alexdavies.net/twitter-sentiment-analysis/>
* <http://www-nlp.stanford.edu/courses/cs224n/2009/fp/3.pdf>