**M S Ramaiah Institute of Technology**

(An Autonomous Institute, Affiliated to VTU)

MSR nagar, MSRIT post, Bangalore-54

A Dissertation Report on

Movie Sentiment Analysis

Submitted by

Kush Bavishi 1MS12CS049

Udhav Chandel 1MS12CS123

Nagashree Devayani Urs 1MS12CS064

Pallavi Hemmige 1MS12CS071

*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**

### DECLARATION

I hereby declare that the project entitled “**Movie Sentiment Analysis**” submitted for the "**Project Based Learning**" is our original work and the project has formed the basis for the award of B.E degree in Computer Science.

Signature of the Students:

Place:

Date:

**ACKNOWLEDGEMENT**

We would like to thank our HOD Dr. K.G.Srinivasa(Professor). Without their push and motivation, this project would have not been completed. Their continuous support and help made this project feasible. Without his support we would not have been able to excel in our project. Lastly, we would like to thanks all our friends and teammates for believing in each other and our capabilities.

# Abstract

This paper presents a study of supervised learning and machine learning techniques in classifying movie reviews by semantic meaning. We use movie reviews from Rotten Tomatoes as our data set and classify text by subjectivity/objectivity and negative/positive attitude. We propose different approaches in extracting text features such as bag-of-words model, using large movie reviews corpus, restricting to adjectives and adverbs, handling negations, bounding word frequencies by a threshold. We evaluate their effect on accuracy by using a machine learning method- Naive Bayes. We reduce our test dataset using Map-Reduce technique on the Hadoop framework.

We conclude our study with, the sentiments found within comments, feedback or critiques to provide useful indicators for many different purposes. These sentiments can be categorized either into five categories: negative, somewhat negative, neutral, somewhat positive, and positive. In this respect, a sentiment analysis task can be interpreted as a classification task where each category represents a sentiment.

#### Contents

***Declaration i***

***Acknowledgements ii***

***Abstract iii***

1 **INTRODUCTION 7**

* 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 9**
   1. Software Process Models
   2. Roles and Responsibilities
2. **LITERATURE SURVEY 11**

3.1….Introduction

3.2…Main Body

3.3 Conclusion of Survey

1. **SOFTWARE REQUIREMENT SPECIFICATIONS 13**

4.1 External Interface Requirements

4.1.1 User Interfaces

4.1.2 Hardware Interfaces

4.1.3 Software Interfaces

4.1.4 Communication Interfaces

4.2 Functional Requirements

4.2.1 Functional Requirement

4.3 Software System Attributes

4.3.1 Reliability

4.3.2 Availability

4.3.3 Security

4.3.4 Portability

4.3.5 Maintainability

4.3.6 Performance

4.4 Performance Requirements

4.5 Database Requirement

4.6 Design Constraints

4.7 Other Requirements

1. **DESIGN 17**
   1. Introduction
   2. Architecture Design
   3. Class Diagram and Classes (represent Inheritance, Aggregation and Association)
   4. Sequence Diagram
   5. Data flow diagram
2. **IMPLEMENTATION 22**
   1. Tools Introduction
   2. Technology Introduction
   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 25**

7.1 Results and Snapshots

1. **CONCLUSION & SCOPE FOR FUTURE WORK 29**
2. **REFERENCES 30**

**1. INTRODUCTION**

**1.1 General Introduction**

The ways people express their opinions and sentiments have changed in the past few years thanks to the arrival of social networks, web communities, and other online media. The extraction of knowledge from the large amount of information on the Web can be a key factor to those who want to engrave an idea into the people of the world and thus reaching out to many. The automatic analysis of online opinions requires an understanding of language text. Sentiment analysis aims to benchmark your sentiment analysis ideas on the Rotten Tomatoes dataset. You are asked to label phrases on a scale of five values: negative, somewhat negative, neutral, somewhat positive, and positive. Obstacles like sentence negation, sarcasm, terseness, language ambiguity, are taken into account in the project.

**1.2 Statement of the Problem**

The goal of our project is to analyze textual content from the perspective of the opinions and viewpoints they hold. The dataset is comprised of tab-separated files with phrases from the Rotten Tomatoes dataset. In the train and test data set each phrase has a PhraseId. Each sentence has a SentenceId. Phrases that are repeated are only included once in the data. (a) train.tsv contains the phrases and their associated sentiment labels. We have additionally provided a SentenceId so that you can track which phrases belong to a single sentence. (b) test.tsv contains just phrases. You must assign a sentiment label to each phrase. The sentiment labels are that are used in this project are as follows:0 – negative,1 - somewhat negative,2 – neutral,3 - somewhat positive,4 – positive.

**1.3 Objectives of the Project**

The aims are of this project was to construct a database of reviews (words/phrases) on the keywords, utilize machine learning techniques which are based on a training set and will determine the sentimental value of the sentence/word/phrases and to provide a sentimental value after thorough analysis of the movie review data set which can help an individual choose a movie.

**1.4 Project Deliverables**

The sentiment analysis project gives us the ability to gauge the sentiment value of a particular review on the Rotten Tomatoes website. Hence a person will be able to determine the sentiment value of a particular movie.

**1.5 Current Scope**

Much of the current research is focusing on the area of sentiment analysis. People are intended to develop a system that can identify and classify opinion or sentiment as represented in an electronic text. An accurate method for predicting sentiments could enable us, to extract opinions from the internet and predict online customer’s preferences, which could prove valuable for economic or marketing research.

**1.6 Future Scope**

Future researches can be carried out to generate better and fast models for higher order n-grams. Hence movie sentiment analysis can be improved in the future.

**2. PROJECT ORGANIZATION**

**2.1 Software Process Models**

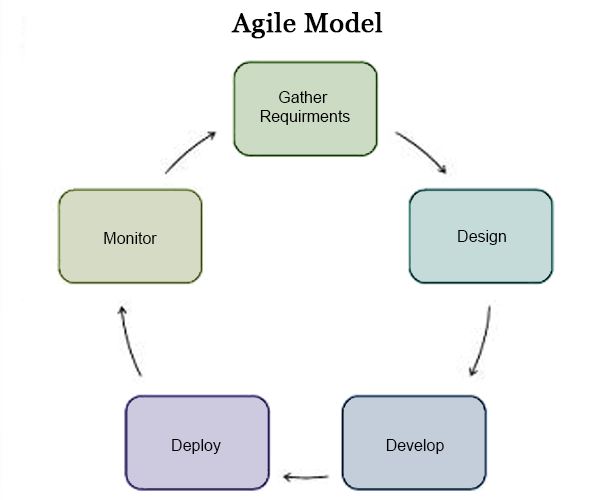


Figure-1: Agile Software Process Model

Agile Software Development is a set of [software development methods](https://en.wikipedia.org/wiki/Software_development_methodologies) in which requirements and solutions evolve through collaboration between self-organizing,[cross-functional teams](https://en.wikipedia.org/wiki/Cross-functional_team). It promotes adaptive planning, evolutionary development, early delivery, continuous improvement, and encourages rapid and flexible response to change.

Agile development model is also a type of [Incremental model](http://istqbexamcertification.com/what-is-incremental-model-advantages-disadvantages-and-when-to-use-it/). Software is developed in incremental, rapid cycles. This results in small incremental releases with each release building on previous functionality. Each release is thoroughly [tested](http://istqbexamcertification.com/why-is-testing-necessary/) to ensure [software quality](http://istqbexamcertification.com/what-is-software-quality/) is maintained. It is used for time critical applications.  Extreme Programming (XP) is currently one of the most well-known agile [development life cycle model](http://istqbexamcertification.com/what-are-the-software-development-models/). The diagram for agile development process is as shown above.

**2.2 ROLES AND RESPONSIBILITIES**

1. Kush Bavishi was assigned the task of writing the code for extracting raw data, to parse it and to obtain the reducing data using hadoop, and also formed the software engineering model.

2. Udhav Chandel was assigned the task of writing the code in eclipse for entering train data into the mySQL database, and also performed the visualization of data using SAP Lumira.

3. Pallavi Hemmige was responsible for comparing and testing the train data with the test data, and also prepared the white paper.

4.Nagashree Devayani Urs was responsible for obtaining the final sentiment value and classifying the data into the respective categories according to sentiment value and preparing the ieee paper.

**3. LITERATURE SURVEY**

**3.1 Introduction**

Lina Zhou et al., investigated movie review mining using machine learning and semantic orientation. Supervised classification and text classification techniques are used in the proposed machine learning approach to classify the movie review. A corpus is formed to represent the data in the documents and all the classifiers are trained using this corpus. Thus, the proposed technique is more efficient. Though, the machine learning approach uses supervised learning, the proposed semantic orientation approach uses “unsupervised learning” because it does not require prior training in order to mine the data. Experimental results showed that the supervised approach achieved 84.49% accuracy in three-fold cross validation and 66.27% accuracy on hold-out samples. The proposed semantic orientation approach achieved 77% accuracy of movie reviews. Thus, the study concludes that the supervised machine learning is more efficient but requires a considerable amount of time to train the model. On the other hand, the semantic orientation approach is slightly less accurate but is more efficient to use in real time applications. The results confirm that it is practicable to automatically mine opinions from unstructured data.

**3.2 Main Body**

Bo Pang et al., used machine learning techniques to investigate the effectiveness of classification of documents by overall sentiment. Experiments demonstrated that the machine International Journal of Computer Applications (0975 – 888) Volume 47– No.11, June 2012 37 learning techniques are better than human produced baseline for sentiment analysis on movie review data. The experimental setup consists of movie-review corpus with randomly selected 700 positive sentiment and 700 negative sentiment reviews. Features based on unigrams and bigrams are used for classification. Learning methods Naïve Bayes, maximum entropy classification and support vector machines were employed. Inferences made by Pang et al., is that machine learning techniques are better than human baselines for sentiment classification. Whereas the accuracy achieved in sentiment classification is much lower when compared to topic based categorization. Zhu et al., proposed aspect based opinion polling from free form textual customers reviews. The aspect related terms used for aspect identification was learnt using a multi-aspect bootstrapping method. A proposed aspect-based segmentation model, segments the multi aspect sentence into single aspect units which was used for opinion polling. Using a opinion polling algorithm, they tested on real Chinese restaurant reviews achieving 75.5 percent accuracy in aspect-based opinion polling tasks. This method is easy to implement and are applicable to other domains like product or movie reviews. Jeonghee Yi et al., proposed a Sentiment Analyzer to extract opinions about a subject from online data documents. Sentiment analyzer uses natural language processing techniques. The Sentiment analyzer finds out all the references on the subject and sentiment polarity of each reference is determined. The sentiment analysis conducted by the researchers utilized the sentiment lexicon and sentiment pattern database for extraction and association purposes. Online product review articles for digital camera and music were analyzed using the system with good results.

**3.3 Conclusion of Survey**

In this literature survey paper it is seen that sentiment analysis/opinion mining play vital role to make decision about product /services. Opinion mining not only encompasses concepts of text mining but also the concepts of information retrieval. Major challenges in opinion mining includes feature weighting which plays a crucial role for good classification. Also it is seen that soft computing techniques have not been extensively used in the literature. Without opinion life is like an empty vessel. The work can be further extended to emerging areas like Mobile learning and investigation with soft computing techniques like neural network.

**4. Software Requirement Specifications**

**4.1 External Interface Requirements**

4.1.1 User Interface:

The user interface we are using for our application involves the project’s output being displayed on the console.

The output containing the sentimental analysis for the movie will be displayed on the screen based on the data that has been provided.

4.1.2 Hardware Interface:

An HP laptop with an AMD A8 Quad core processor (1.9 Ghz) and 4GB of RAM was used.

4.1.3 Software Interface:

* The Ubuntu 14.04 is used along with Hadoop (2.7.1 ) for Map-Reduce, the integrating of test and train data.
* Obtaining the sentimental analysis is done by eclipse indigo using JDBC.
* MySQL as the database and XAMPP servers are used for the application.
* For the data gathering rotten tomatoes is the source of data.

4.1.4 Communication Interface:

There are no communication interfaces used in this project, since analysis is done on the remote system.

**4.2 Functional Requirements**

4.2.1 Retrieving Input:

The inputs retrieved are datasets of test and train data obtained from Rotten Tomatoes. Any user can provide his movie analysis data which is stored into the Hadoop architecture. The system will reduce these datasets using various map reduce techniques.

4.2.2 Processing of data:

Hadoop will take input of datasets, which will be processed and reduced using Phrase Count Map-Reduce algorithm. The Train data will be extracted and entered into a MySQL database.

4.2.3 Sentimental Analysis:

After integrating the train data with the database, JDBC is used for obtaining the sentimental analysis of each phrase in the test data. The sentimental value is calculated by checking the phrases with reference to the train data. After the analysis the sentimental rating is obtained. There are five sentimental values:

- Negative = 0

- Somewhat negative = 1

- Neutral = 2

- Somewhat positive = 3

- Positive = 4

4.2.4 Output:

The output showing the overall sentimental value of the data is displayed on the console. This must be clear and specific by providing a value based on the list of sentimental values.

**4.3 Software system attributes**

4.3.1 Reliability: The software will meet all of the functional requirements without any unexpected behavior. At no time should the gauge output display incorrect or outdated information without alerting the user to potential errors.

4.3.2 Availability: The software reviewing the data will be available at all times, as long as the device we are using is in proper working order. The functionality of the software will depend on the different movie data sets we use.

4.3.3 Security: The software will be totally secure without any chances of data getting corrupted or the data being lost. The only security breach which can take place is if the system containing the application is stolen.

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

4.3.5 Portability: This software will run on any Linux based system which supports Hadoop and Java.

4.3.6 Performance: The system must be interactive and the delays involved must be less. So in every action-response of the system, there are no immediate delays. The delay in the system increases with the amount of data sets which we are to be read and processed.

**4.4 Performance Requirements**

a. Real-Time: The software will provide up-to-date information, limited only by the rate of input of movie data. The output should display the results at all times, for that particular movie review.

b. System Resource Consumption: Resource consumption of this software should not reach an amount that the device becomes unusable. The software should be capable of handling a large amount of data.

When you are building a sentiment analysis system, you should first split your data set into 2 sets. 

1. Training Set (around 60%) - This set will be used for training your classifier.
2. Test Set (around 40%) - You would not use this test set until you have finalized your model and is ready for production. You would then run your system on this set to check the performance of your system i.e. whether your model generalized properly or not.

**4.5 Database Requirement**

The reviews taken from Rotten tomatoes will be stored on an excel spreadsheet. Excel is an excellent program for storing large amounts of data as well as being easy to upload the data .The data will have five columns, column four will have the score of the reviews ( positive, negative, very positive and very negative), column three will store the actual content. Each row will represent an individual phrase/word of the review.

The attributes of the database are:

1. Phrase ID
2. Sentence Number
3. Phrase
4. Sentiment value

**4.6 Design Constraints**

Rotten Tomatoes API has some limitations such as Rotten Tomatoes API can only return a fixed maximum amount of reviews (1500). The return of a maximum number of reviews may not be met sometimes as there are not enough reviews for the particular keyword.

Rotten Tomatoes Data: The application must comply with the Rotten Tomatoes Developer terms of service. This includes the following: Defining an application privacy policy (what we do with reviews, user data, etc.) , Not redistributing Movie Reviews , providing a link to Rotten Tomatoes sign-up if user does not have a registered account.

**5. DESIGN**

**5.1 Introduction**

* 1. Number of Modules

In the movie sentiment analysis we use 4 modules. They are:

1. Extract the phrases from the test data

2. Map-reduce operations performed on the test data

3. Putting Train data into the database

4. Calculating the sentimental value

* 1. Modules description
     1. Extract the phrases from the test data: Our test data has several extra columns from which we are extracting the required data and writing it onto a file
     2. Map-Reduce operations performed on the test data: Our test data will be reduced by implementing the map-reduce operation on phrases using Hadoop to give us the count of non-unique phrases.
     3. Putting Train data into database: The train data will be extracted and entered into the database using JDBC concepts.
     4. Calculating the sentimental value: Each of the phrases from the output of the reduced test data will be checked against the train data and the final sentimental value will be calculated.
  2. Algorithm design

Naive Bayes Algorithm : It is a popular method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the feature.

Bag Of Words: : The bag-of-words model is commonly used in methods of document classification, where the (frequency of) occurrence of each word is used as a feature for training a classifier.

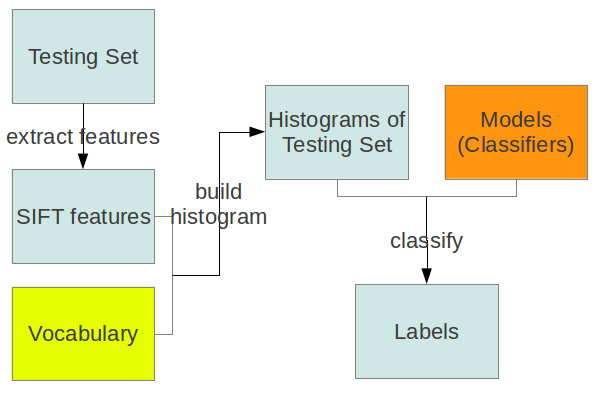


Figure 2: Bag of Words Model

Ngram.: In the fields of computational linguistics and probability, an ***n*-gram** is a contiguous sequence of *n* items from a given sequence of text or speech. The items can be phonemes, letters, words or base pairs according to the application. The *n*-grams typically are collected from a text or speech corpus.

An *n*-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram" (or, less commonly, a "digram"); size 3 is a "trigram". Larger sizes are sometimes referred to by the value of*n*, e.g., "four-gram", "five-gram", and so on.

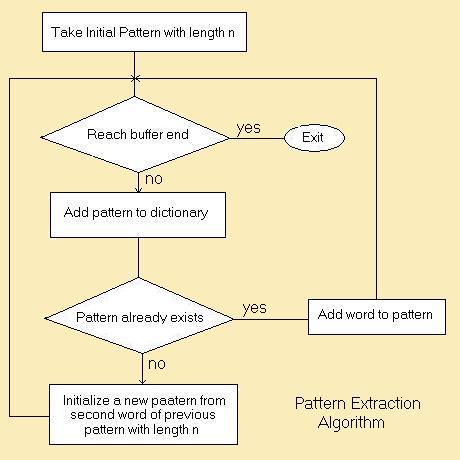


Figure 3: N-gram Model

**5.2 Architecture Design**

The architectural design is the design of the entire software system; it gives a high-level overview of the software system, it provides information on the decomposition of the system into modules (classes), dependencies between modules, hierarchy and partitioning of the software modules.

****

Figure 4: Architecture Design

**5.3 Class Diagram**

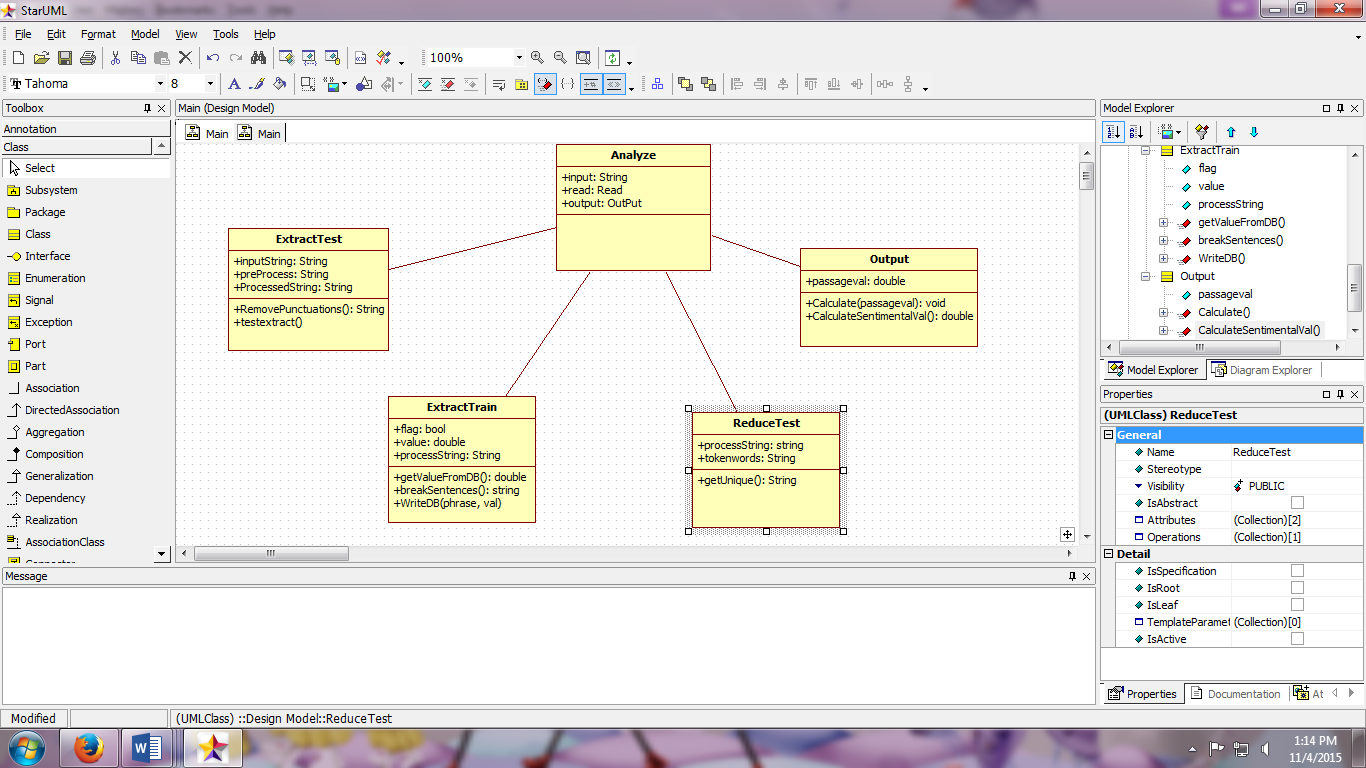
****

Figure 5: Class Diagram of Movie Sentiment Analysis

**5.4 Sequence Diagram**

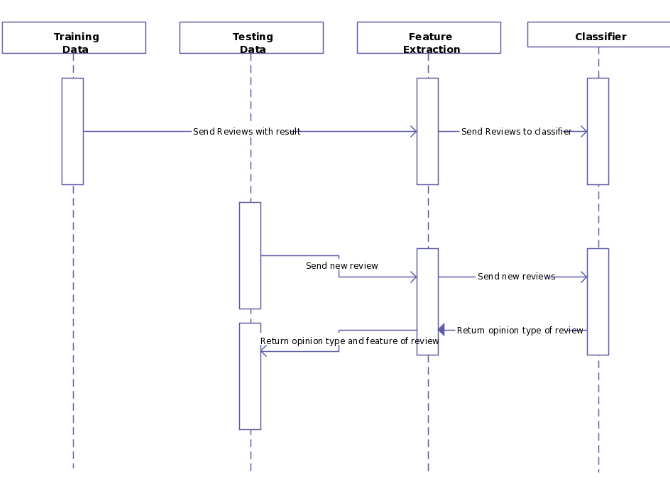
****

Figure 6: Sequence Diagram of Movie Sentiment Analysis

**5.5 Data Flow diagram**

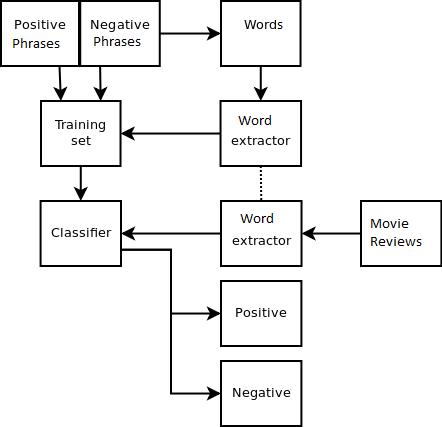
****

Figure 7: Data Flow Diagram of Movie Sentiment Analysis

**6. IMPLEMENTATION**

**6.1 Tools Introduction**

Hadoop framework:

Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines or racks of machines) are commonplace and thus should be automatically handled in software by the framework.

SAP Lumira:

SAP Lumira (formerly called SAP Visual Intelligence) is a self-service, [data visualization](http://searchbusinessanalytics.techtarget.com/definition/data-visualization) application for business users. SAP Lumira grew out of [SAP Business Explorer](http://searchsap.techtarget.com/definition/BEx), the company's previous, but more basic, self-service data visualization tool. According to SAP, the two are  meant to go hand-in-hand.

MySQL: is an open-source relational database management system.

Xampp server:

The Xampp is a free and open source cross-platform web server solution stack package developed by Apache Friends, consisting mainly of the Apache HTTP Server, MySQL database, and interpreters for scripts written in the PHP and Perl programming languages.

Ubuntu Operating System:

Ubuntu Linux operating system and distribution, with Unity as its default desktop environment for personal computers including smart phones in later versions.

**6.2 Technology Introduction**

Java: (Sun Java 7):

Java is a set of computer software and specifications developed by Sun Microsystems, later acquired by Oracle Corporation that provides a system for developing application software and deploying it in a cross-platform computing environment.

**6.3 Overall view of the project in terms of implementation**

The sentimental value is obtained according to the sentimental analysis scale that we have utilized i.e somewhat negative, negative, positive, and somewhat positive and so on. Hadoop is the framework that helps us to reduce any redundancy and a java program was written with the help of algorithms such as Naïve Bayes which acts a classifier. Thus the sentimental value of the movie review is obtained.

**6.4 Explanation of Algorithm and how it is been implemented**

Naive Bayes Algorithm : It is a popular method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the feature.

Bag Of Words: : The bag-of-words model is commonly used in methods of document classification, where the (frequency of) occurrence of each word is used as a feature for training a classifier.

Ngram.: In the fields of computational linguistics and probability, an *n*-gram is a contiguous sequence of *n* items from a given sequence of text or speech. The items can be phonemes, letters, words or base pairs according to the application. The *n*-grams typically are collected from a text or speech corpus.

An *n*-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram" (or, less commonly, a "digram"); size 3 is a "trigram". Larger sizes are sometimes referred to by the value of*n*, e.g., "four-gram", "five-gram", and so on.

**6.5 Information about the implementation of Modules**

Number of Modules:In the movie sentiment analysis we use 4 modules. They are:

1. Extract the phrases from the test data

2. Map-reduce operations performed on the test data

3. Putting Train data into the database

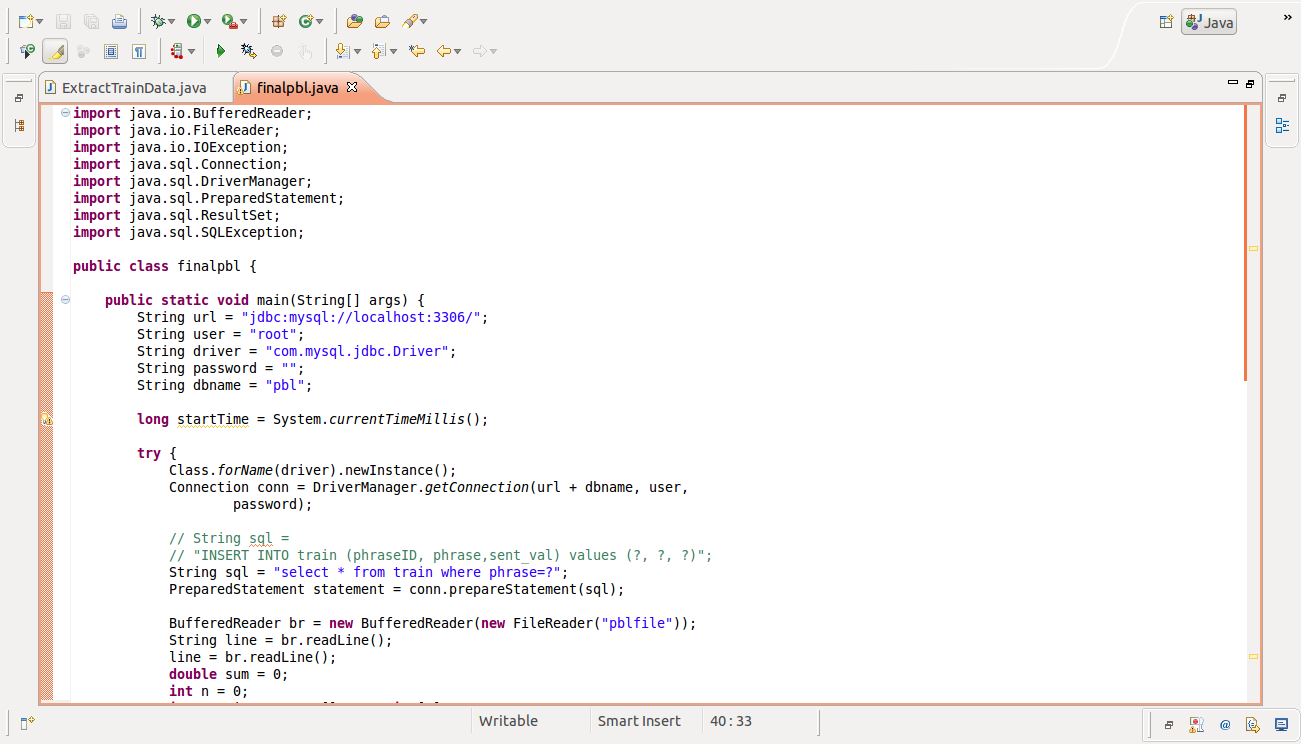
4. Calculating the sentimental value

Modules description

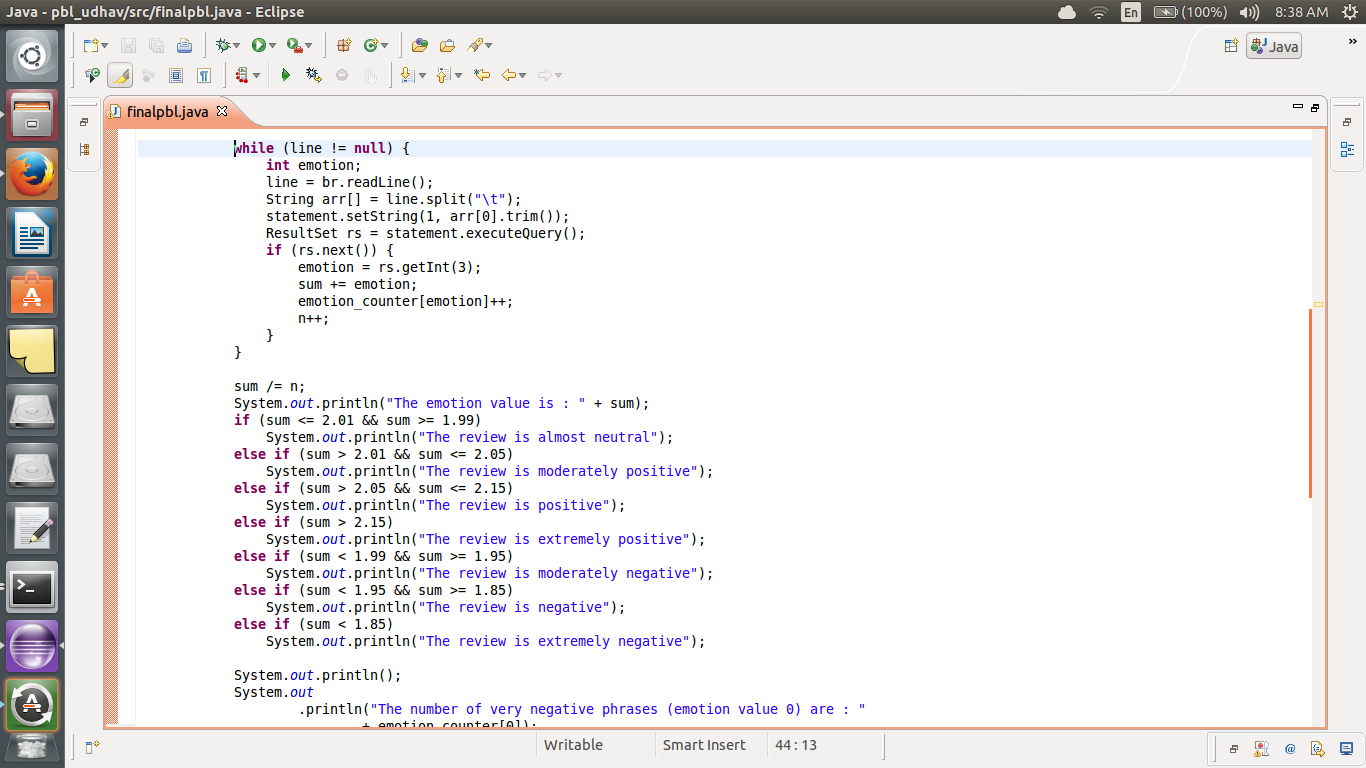
* Extract the phrases from the test data: Our test data has several extra columns from which we are extracting the required data and writing it onto a file
* Map-Reduce operations performed on the test data: Our test data will be reduced by implementing the map-reduce operation on phrases using Hadoop to give us the count of non-unique phrases.
* Putting Train data into database: The train data will be extracted and entered into the database using JDBC concepts.
* Calculating the sentimental value: Each of the phrases from the output of the reduced test data will be checked against the train data and the final sentimental value will be calculated.

**7. TESTING**

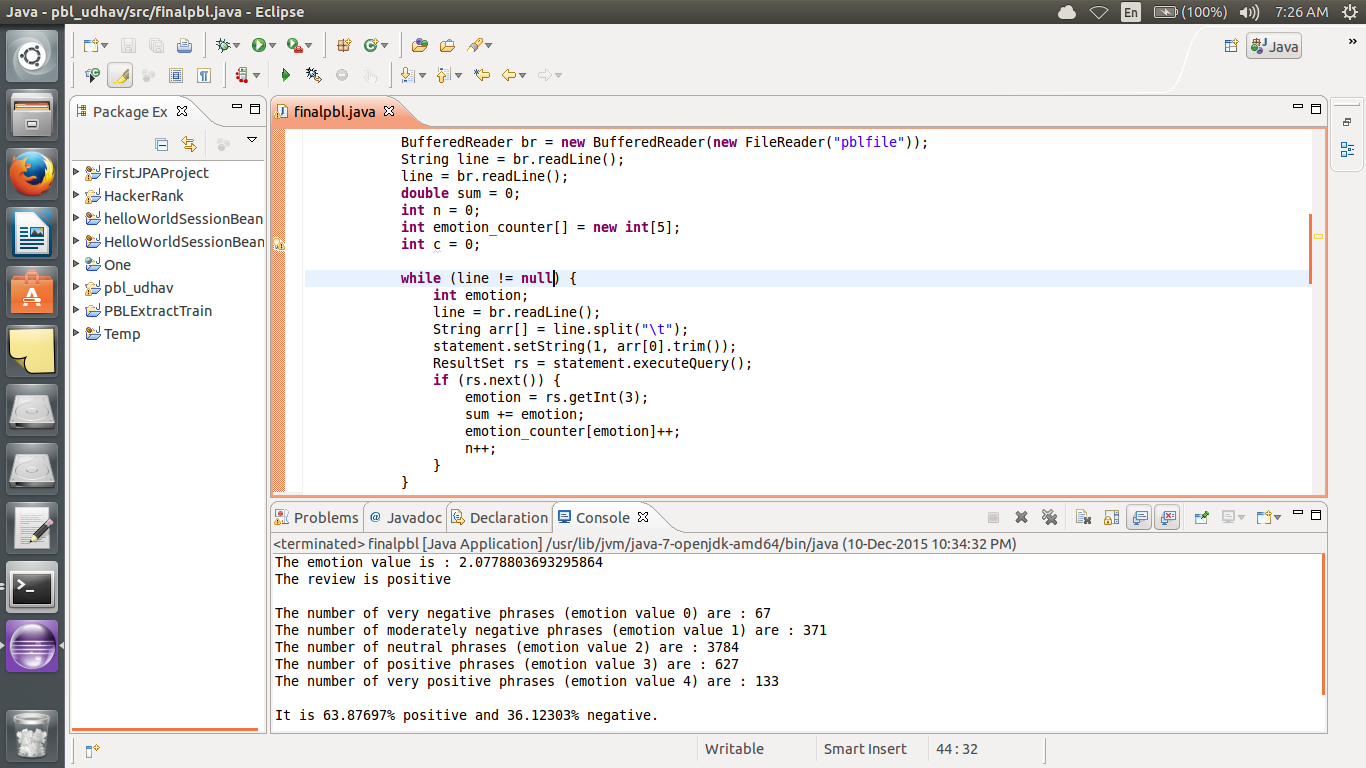
**7.1 Results and Snapshots**



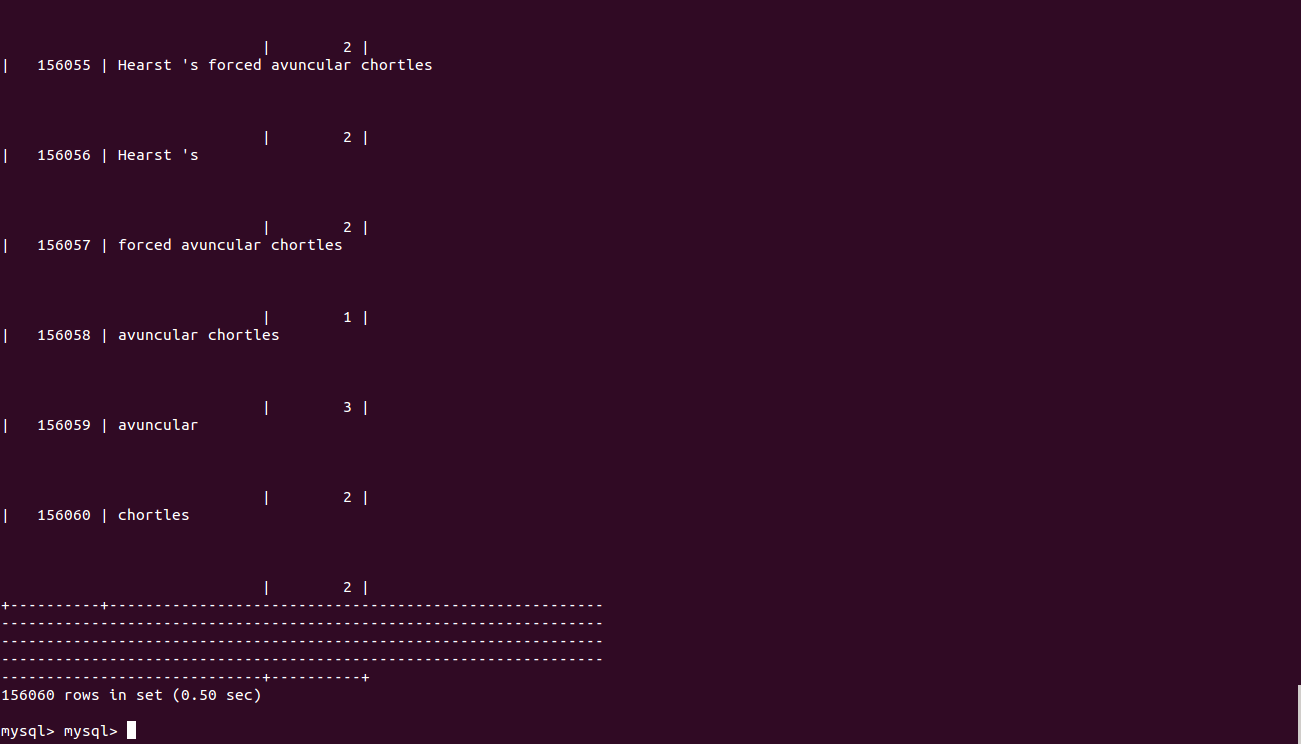
Snapshot-1: First Part of Program for Sentiment Value

****

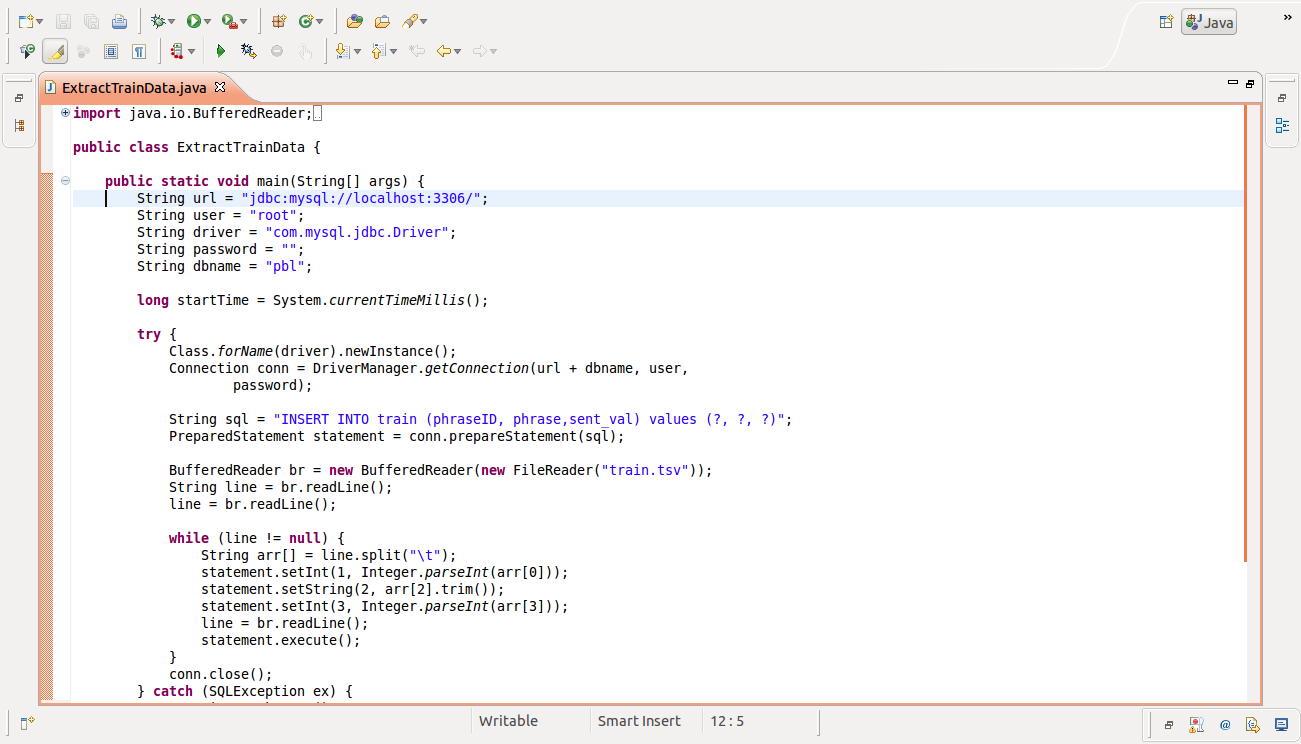
Snapshot-2: Second Part of Program for Sentiment Value

****

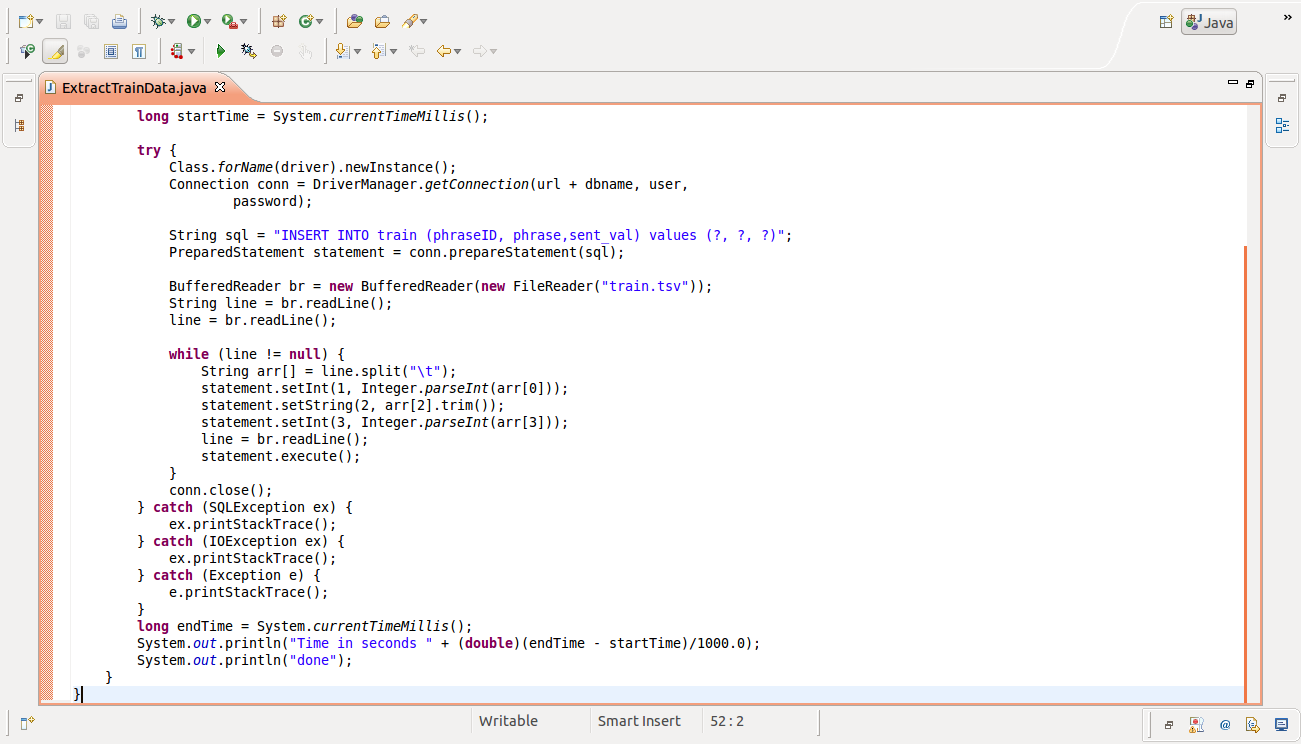
Snapshot-3: Output showing Sentiment Value



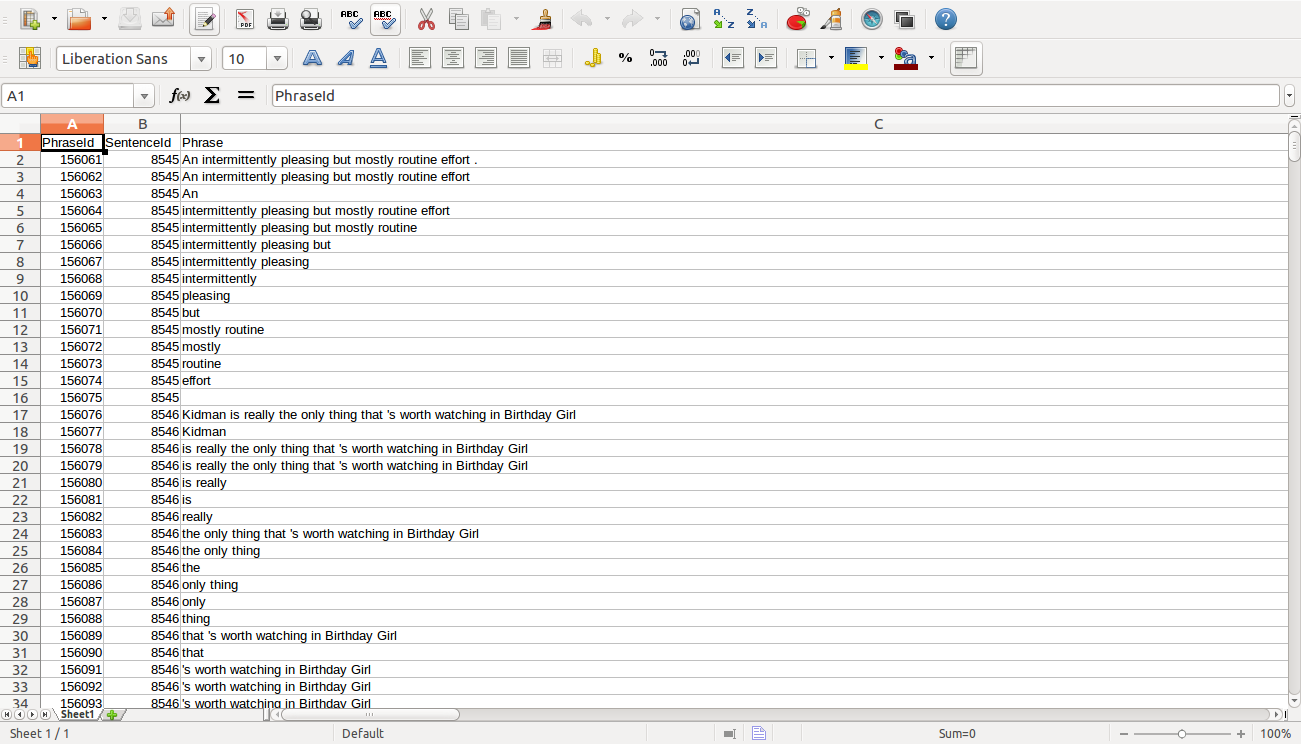
Snapshot-4: Ubuntu MySQL database for test data



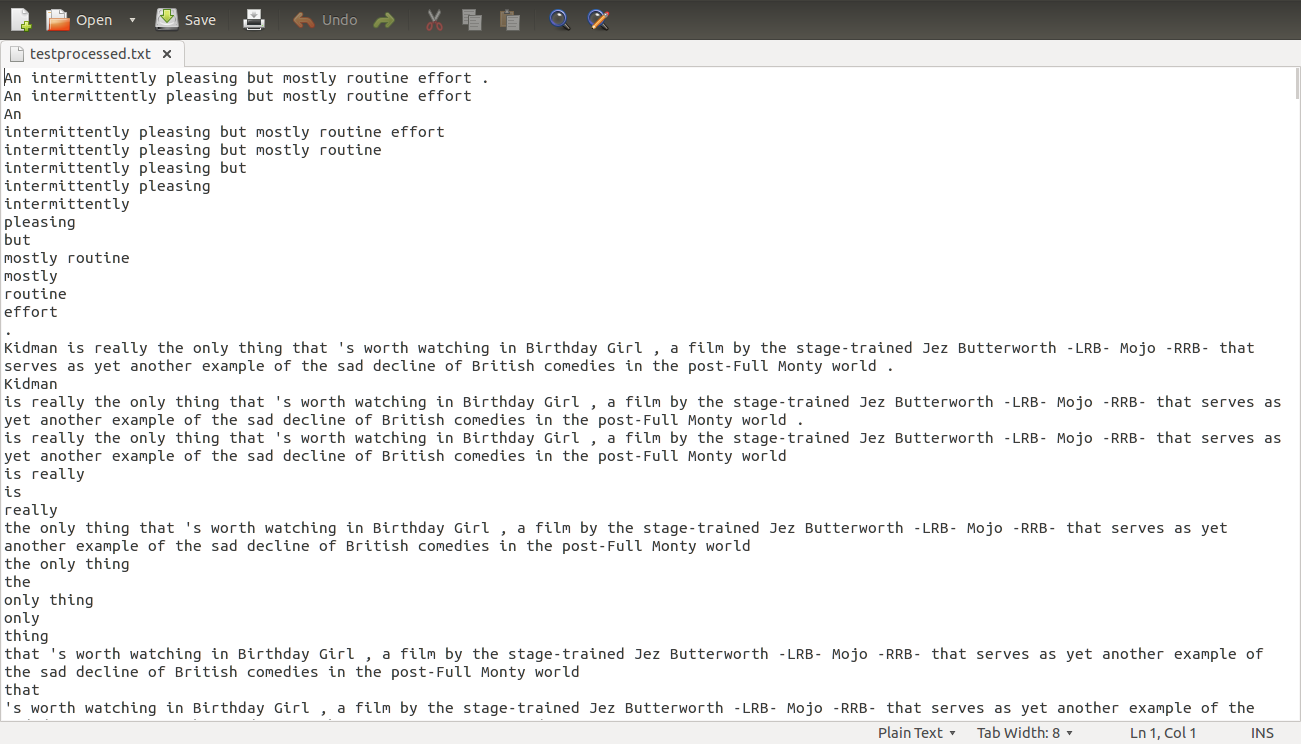
Snapshot-5: Part One of storing data into database



Snapshot-6: Part Two of storing data into database



Snapshot-7: Train data



Snapshot-8: Test Data

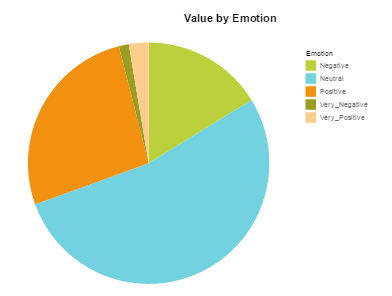


Figure-8: Categorization using SAP Lumira

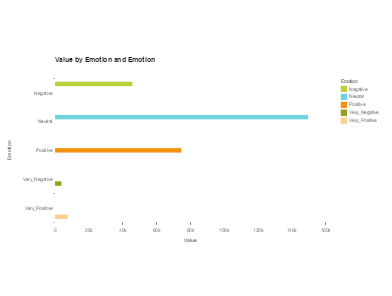


Figure-9: Sentiment value using bar charts

**8. CONCLUSION & SCOPE FOR FUTURE WORK**

In conclusion, the sentiment analysis project gives us the ability to guage the sentiment value of a particular review on the Rotten Tomatoes website. Hence a person will be able to determine the sentiment value of a particular movie. It not only helps assess the viewer the movie he would want to watch based on its sentiment, but also provides an overall categorization of the critics review based on his sentiment of the view. It helps us get a clear understanding of the application of hadoop, the usage of machine learning algorithms and finally visualization of data by SAP Lumira. The sentiment analysis is the future of analytic research, and the analysis on movie reviews is just about to get bigger.

**9. REFERENCES**

[1] Amazon online retailer web site. http://www.amazon.com.

[2] Digg social networking site. <http://www.digg.com>.

[3] Natural language toolkit. <http://www.nltk.org>.

[4] Rottentomatoes movie review site. <http://www.rottentomatoes.com>.

[5] Twitter social networking site. <http://www.twitter.com>.

[6]http://www.cisco.com/web/offer/email/43468/5/Cloud\_Computing\_in\_Higher\_Education.pdf : Cloud Computing in Higher Education: A Guide to Evaluation and Adoption

[7] A. Bagga and B. Baldwin. Algorithms for scoring coreference chains. In In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference, pages 563– 566, 1998.