Sentiment Analysis

Minor Project Report

SUBMITTED IN PARTIAL FULFILLMENT REQUIREMENT FOR

THE AWARD OF DEGREE OF

Bachelor of Technology

(Computer Science & Engineering)

SUBMITTED BY

Amulya Garg, Nikant, Prabhudeep

(UNIVERSITY ROLL No. 1410811,1410891,1410899)

April, 2017

PUNJAB TECHNICAL UNIVERSITY

JALANDHAR (PUNJAB), INDIA

ABSTRACT

Sentiment analysis or opinion mining is one of the major tasks of NLP (Natural Language Processing). Sentiment analysis has gain much attention in recent years. In this paper, we aim to tackle the problem of sentiment polarity categorization, which is one of the fundamental problems of sentiment analysis. A general process for sentiment polarity categorization is proposed with detailed process descriptions. Data used in this study are online product reviews collected from any online resource using BeautifulSoup. Experiments for both sentence-level categorization and review-level categorization are performed with promising outcomes. At last, we also give insight into our future work on sentiment analysis

ACKNOWLEDGEMENT

I am highly grateful to **Dr. M. S. Saini**, Director, Guru Nanak Dev Engineering College, Ludhiana, for providing opportunity to carry out Minor project from Janurary-May 2017.

Er. Sukhjit Singh Sehra has provided great help in carrying out the my work and is acknowledged with reverential thanks. Without the wise counsel and able guidance, it would have been impossible to complete the thesis in this manner.

I would like to express thanks profusely to **Dr. Parminder Singh**, Professor and Head (Computer Science & Engineering) for stimulating me time to time. I would also like to thank **Er. Sukhjit Singh Sehra**, Training and Placement coordinator (Computer Science and Engineering) and to entire faculty, staff of computer science and engineering. I also thanks my friends who devoted their valuable time and helped me in all possible ways towards successful completion of this work.

(Amulya Garg, Nikant, Prabhudeep)

LIST OF FIGURES

LIST OF TABLES

Table No.	Table Description	Page No.

LIST OF ABBREVIATIONS

Abbreviation	Full Form
GNDEC	Guru Nanak Dev Engineering College
UI	User Interface
API	Application Programming Interface
FAB	Floating Action Bar

TABLE OF CONTENTS

Chapter 1

Introduction

Introduction to project

Sentiment is an attitude, thought, or judgment prompted by feeling. Sentiment analysis [1-8], which is also known as opinion mining, studies peoples sentiments towards certain entities. Internet is a resourceful place with respect to sentiment information. From a users perspective, people are able to post their own content through various social media, such as forums, micro-blogs, or online social networking sites. From a researchers perspective, many social media sites release their application programming interfaces (APIs), prompting data collection and analysis by researchers and developers.

For instance, Twitter currently has three different versions of APIs available [9], namely the REST API, the Search API, and the Streaming API. With the REST API, developers are able to gather status data and user information; the Search API allows developers to query specific Twitter content, whereas the Streaming API is able to collect Twitter content in realtime. Moreover, developers can mix those APIs to create their own applications. Hence, sentiment analysis seems having a strong fundament with the support of massive online data.

What is Sentiment?

Sentiment = feelings

1.Attitudes

2.Emotions

3.Opinions

Usually reviews are given in text format. Single product has number of reviews. It is hardly possible to read each review in detail. Many researches shown pictorial representation is more effective and can be memorized, understood easily rather than textual representations. So if we are going to convert textual reviews into visual format, it will enhance reliability in decision making.

Project Category

This project is categorized System Development. This project is a product review System, which analyse the product reviews and shows out the sentiment in chart forms. It gives us detailed examination of the product reviews.

Objectives

Sentiment analysis or opinion mining is a field of study that analyzes peoples sentiments, attitudes, or emotions towards certain entities. This paper tackles a fundamental problem of sentiment analysis, sentiment polarity categorization.

1. Online product reviews any website can be selected as data used for this study.

2.A sentiment polarity categorization process has been proposed along with detailed description of each step.

3.Experiments for both sentence-level categorization and review-level categorization have been performed.

Problem Formulation

What other people think has always been an important piece of information for most of us during the decision-making process. The product makers need to analyze the reviews of the product for planning and executing the future projects.

Recognition of Needs

After launching a product, the company needs to analyze the review of the product. Going through each and every product review can become very exhaustive. Thus we need such a system which interactively shows the sentiments of product reviews using graphs, so that we can save our time and efforts.

Existing System

Existing work shows that various approaches are used for Sentiment Analysis like machine learning, corpus based, NLP based or even based on clustering. Also few researches consider neutral review for analysis. Many of them do not have visual representations for end results or complex visual representations which are not user oriented.

Proposed System

While doing analysis of product reviews, company needs to know:

- Does the product has satisfied a big pool of customers?
- Are the customer reviews positive or negative?
- Analyze each individual's reviews
- Finally producing the overview of analyzed results.

Thus keeping these points a system should be developed which represents product reviews in a better way.

Unique features of the System

- Unsupervised Machine learning i.e Machine trains itself on the available data to the machine.

- Analyzing the comment using words, then sentences, then whole paragraph and then giving the over all result.
- Showing the results of the analyzed data.
- Plotting the pie chart of analyzed data.

Chapter 2

Requirement Analysis and System Specification

Feasibility Study

In feasibility study phase we had undergone through various steps:

1.**Technical Feasibility** : This project Sentiment Intelligence will be platform independent. It is coded in Python. Hardware requirements used are compatible with all OS. The system can also be expanded as per the needs of requirement specification.

2.**Economic Feasibility**: The cost required in the proposed system is economic and affordable as no additional hardware is required for the project. We need only manpower i.e skilled person who knows python and has worked on various python libraries like SCIKIT-LEARN, NLTK, BEAUTIFULSOUP etc.

3.**Legal Feasibility**: Since this application involves no specific society membership, it is totally feasible legally. It the proposed system conflicts with legal requirements like data protection acts or social media laws.

4.**Operational Feasibility**: As per requirement, the proposed solution will provide a tool to the various film production companies to analyze their business needs and customer review for their project. 5. **Scheduling Feasibility**: The project is estimated to complete within a time span of 60 working days. 24 days for SRS documentation, 24 days for product design and code, 12 days for unit and integration testing.

Software Requirement Specification

Introduction

Purpose

- To develop an easier system for review analysis of a product
- Clear and Concise results available after analysis
- Business oriented results
- Results can be used for enhancing the upcoming products on the basis of already launched product review.

Scope

- Product has a brighter future. It can be used as a review analysis system for any product launched by any company.
- Artificial Intelligence can be further used in the product to automate it to a particular type of inputs.

General Requirements

Project Perspective

We need people who have knowledge and hands on experience on python and data science libraries like nltk, beautifulsoup, numpy etc. We also need personal systems for every team member with python installed on it and having at least 2gb Ram installed on the system.

Product Function

Product will provide a system to take input of the product user and will analyze its positivity or negativity of the comment.

Constraints

Product needs a csv file as input. So, you need to convert user comments into .csv format to analyze the data generated by the user.

Specific Requirement

- Need a .csv format input.
- Need permission of accessing the survey website legally.
- Need a working Internet connection.
- Need specific manpower who have knowledge of python.

Project Development Plan

Purpose

Produce an Analyzer System to :

- Take user inputs as .csv
- To show positivity and negativity of the comment

Team Members

- **Nikant:** Programmer and Team Lead
- **Amulya Garg:** Programmer and planner
- **Prabhudeep Singh:** Programmer and Milestone writer

Timeline

Days 0-15 – Requirement Specifications

Days 15-22 – Design Prototype

Days 22-25 – Module Design

Days 25-50 – Coding

Days 50-65 – Analyzing the data

Days 65-70 – Generating the results

Days 75-80 – Publishing the results

Days 80-90 – Unit and Integration Testing

Day 90 – Project Fininshed

Expected Hurdles

- Training Machine to generate bag of words
- Fetching output in .csv format
- Plotting Charts

SDLC Model: Rapid Application Development (RAD)

Introduction to RAD

RAD model is Rapid Application Development model. It is a type of incremental model. In RAD model the components or functions are developed in parallel as if they were mini projects. The developments are time boxed, delivered and then assembled into a working prototype. This can quickly give the customer something to see and use and to provide feedback regarding the delivery and their requirements.

Advantages of RAD

- Reduced development time.
- Increases reusability of components
- Quick initial reviews occur
- Encourages customer feedback
- Integration from very beginning solves a lot of integration issues.

Why RAD Model?

- RAD should be used when there is a need to create a system that can be modularized in 2-3 months of time.
- It should be used if theres high availability of designers for modeling and the budget is high enough to afford their cost along with the cost of automated code generating tools.
- RAD SDLC model should be chosen only if resources with high business knowledge are available and there is a need to produce the system in a short span of time (2-3 months).

Chapter 3

System Design

Design Approach

Design Approach for this system is Object Oriented. The key ideas of the object oriented approach are :

- Objects
- Encapsulation
- Class and Inheritance
- Instances and Instantiation
- Methods and Messages

On of the main principles in the object oriented (OO) approach is that of abstraction, not of data structures and processes separately but both together. An object is a set of data structures and the methods or operations needed to access those structures.

Encapsulation of data structures and methods means that only the methods associated with the object can access the internal data structures. An object is a packaged item of information including the processes which manipulate it. Although we saw above that it is possible to create such packages in e.g. C, the language did not specifically support or enforce the encapsulation.

The objects in OO software are reusable components which may often be used in different applications. In OO environments, object libraries may be available for the programmer to build into solutions.

Generic objects representing classes can also be defined and objects representing subclasses can inherit the data structures and methods from the broader classes to which they belong. A particular instance of a class is represented by instantiating the data structure to the appropriate values.

A programmer may define the structure of an object to represent person and then a new object for employee. The employee object will inherit the data structures and methods from the person object. Note that this inheritance property is equivalent to the is a relation in semantic nets.

Detail Design

Input

The input to the system is a .csv file, in which different polarities of different id's are defined. A .csv file is a comma seperated values integrated in a single file.

Machine Learning

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can change when exposed to new data.

The process of machine learning is similar to that of data mining. Both systems search through data to look for patterns. However, instead of extracting data for human comprehension – as is the case in data mining applications – machine learning uses that data to detect patterns in data and adjust program actions accordingly. Machine learning algorithms are often categorized as being supervised or unsupervized. Supervised algorithms can apply what has been learned in the past to new data. Unsupervised algorithms can draw inferences from datasets.

Machine review analysis

After Machine learning is completed, the machine analys and gives Machine analysis review. The generated Machine analysis review from bag of words is further written in a .csv format. Now this .csv file is used to give visualised results of the analysis.

Flowchart

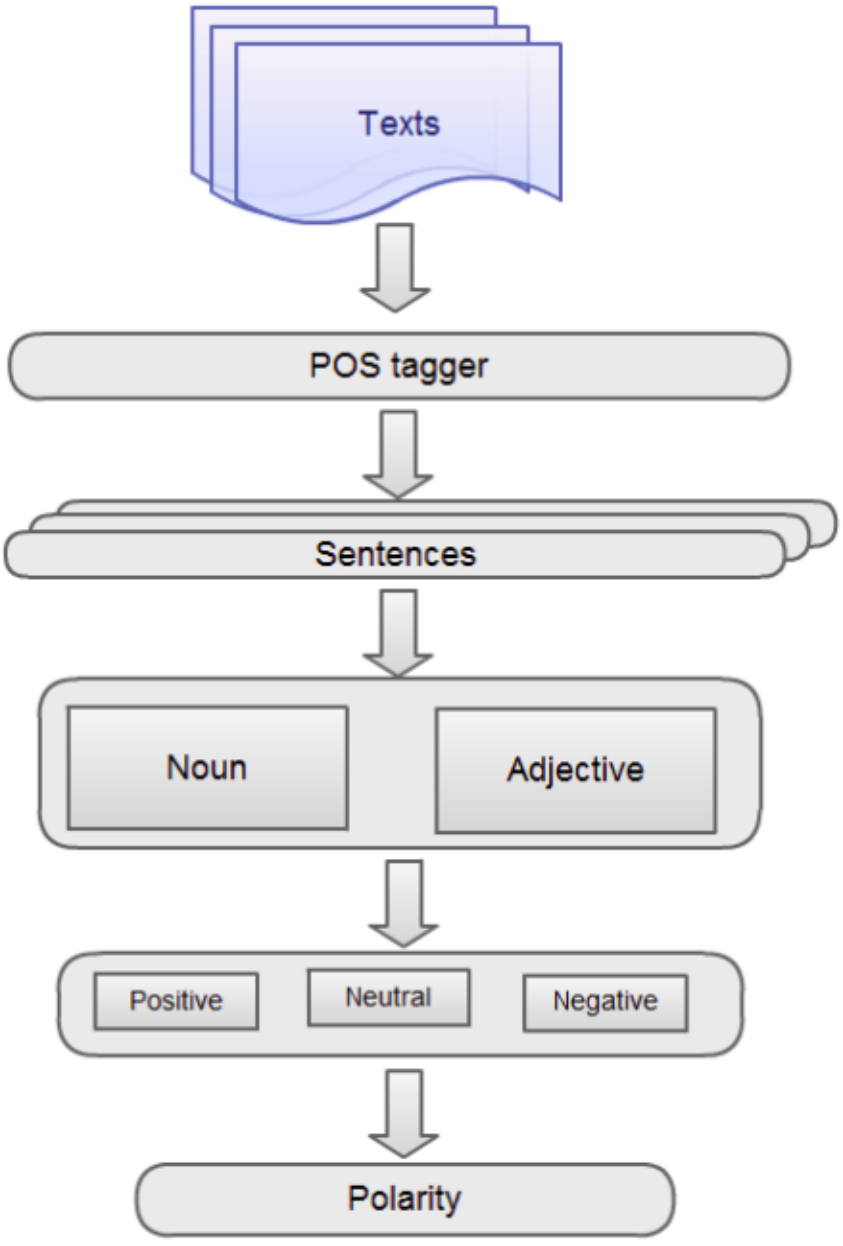

FIGURE 3.1: Flow Chart demonstrating the project

Methodology

Data Preprocessing

Data Preprocessing includes proper fragmentation of data and leaning of data. Here, in research work we are going to use NLP preprocessing techniques like removal of stop words, chunking data, stemming etc. Data preprocessing will lead us to robust data which has less moice. For data preprocessing, use of NLTK library implemented in python is considered. NLTK is a platform for natural language processing developed in python.

POS tagging

POS tagging assists us to identify actual part of sentence which has expression or feelings. It deals with Word Sense Disambiguity(WSD). Single word may have one or more tags. Thus, POS tagging is used to determine single tag for instance of word.

Text Classification

The process of text classification is divided into two stages: Training stage and Testing stage. In training stage, classification model is created using testing dataset. In testing stage, accuracy of classification is evaluated using classification module.

Chapter 4

Implementation, Testing and Maintanence

Introduction to language

In the back end,Python is used. Concepts and libraries related to data science and deep learning are used. Python is one of the world's most important and widely used computer languages, and it has held this distinction for many years. Unlike some other computer languages whose influence has weared with passage of time, while Python has grown.

As of 2017, Python is one of the most popular programming languages in use, particularly for data science applications, with a reported 11 million developers using and working on it.

Applications of Python:

Python is widely used in every corner of world and of human life. Python is not only used in softwares but is also widely used in designing hardware controlling software components. There are more than 100 million python development tool downloads each year.

Following are some other usage of Python :

- Web Applications like Linkedin.com, Snapdeal.com etc.
- Cross platform Mobile application development.
- Internet of things is widely using it.
- Embedded Systems.
- Robotics and games etc.

Supporting Languages

Follwing Libraries are used:

- NLTK(natural language processing toolkit).
- Gensim.
- Scipy.
- bs4.
- Pandas.

Along with these Libraries, some inbuilt methods of python development kit are also used.

Coding Standards

- Use single quotes
- Imports
- camelCase
- PascalCase
- PEP 257 (Docstring Conventions)
- Referencing other code objects with :py:

Project Scheduling

PERT chart

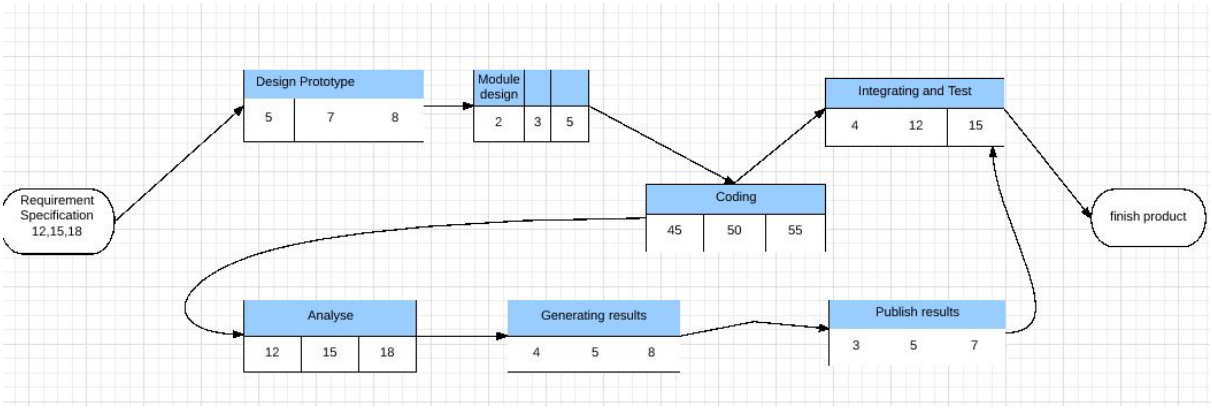

FIGURE 4.1: Pert Chart for project Scheduling

A PERT chart is a project management tool used to schedule, organize, and coordinate tasks within a project. PERT stands for Program Evaluation Review Technique, a methodology developed by the U.S. Navy in the 1950s to manage the Polaris submarine missile program.

A PERT chart presents a graphic illustration of a project as a network diagram consisting of numbered nodes (either circles or rectangles) representing events, or milestones in the project linked by labelled vectors (directional lines) representing tasks in the project. The direction of the arrows on the lines indicates the sequence of tasks.

The PERT chart is sometimes preferred over the Gantt chart, another popular project management charting method, because it clearly illustrates task dependencies. On the other hand, the PERT chart can be much more difficult to interpret, especially on complex projects.

Chapter 5

Results and Discussions

Various Modules of the System

Input

The input to the system is a .csv file, in which different polarities of different id's are defined. A .csv file is a comma seperated values integrated in a single file.

Machine Learning

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can change when exposed to new data.

The process of machine learning is similar to that of data mining. Both systems search through data to look for patterns. However, instead of extracting data for human comprehension – as is the case in data mining applications – machine learning uses that data to detect patterns in data and adjust program actions accordingly. Machine learning algorithms are often categorized as being supervised or unsupervized. Supervised algorithms can apply what has been learned in the past to new data. Unsupervised algorithms can draw inferences from datasets.

Machine review analysis

After Machine learning is completed, the machine analys and gives Machine analysis review. The generated Machine analysis review from bag of words is further written in a .csv format. Now this .csv file is used to give visualised results of the analysis.

Snapshots of System

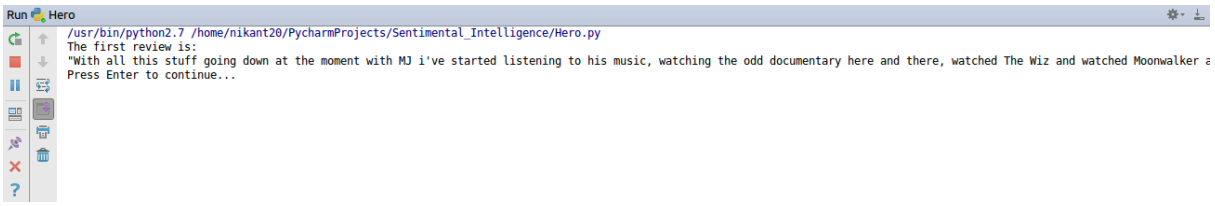

FIGURE 5.1: Fetching the review

In this picture, the reviews from the .csv file is being fetched and is getting ready to be trained.

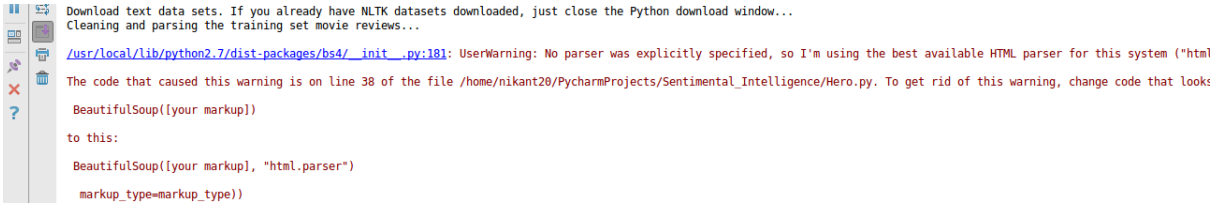

FIGURE 5.2: Downloading the NLTK library

In this, Python downloads the data associated with the NLTK library. If it is already available, then it uses that available data.

FIGURE 5.3: Training the Machine to learn from data

In this, Machine is training itself on the random data from .csv file using random forest algorithm.

In this, after learning from the data, the machine has analyzed the sample data and has predicted the number of Positive and Negative reviews available in the data. Where 0

```
Predicting test labels...

Wrote results to Bag_of_Words_model.csv
0    12782
1    12218
Name: sentiment, dtype: int64
```

FIGURE 5.4: Displaying the results after analyzing the .csv file

stands for Negative value and 1 stands for Positive value.

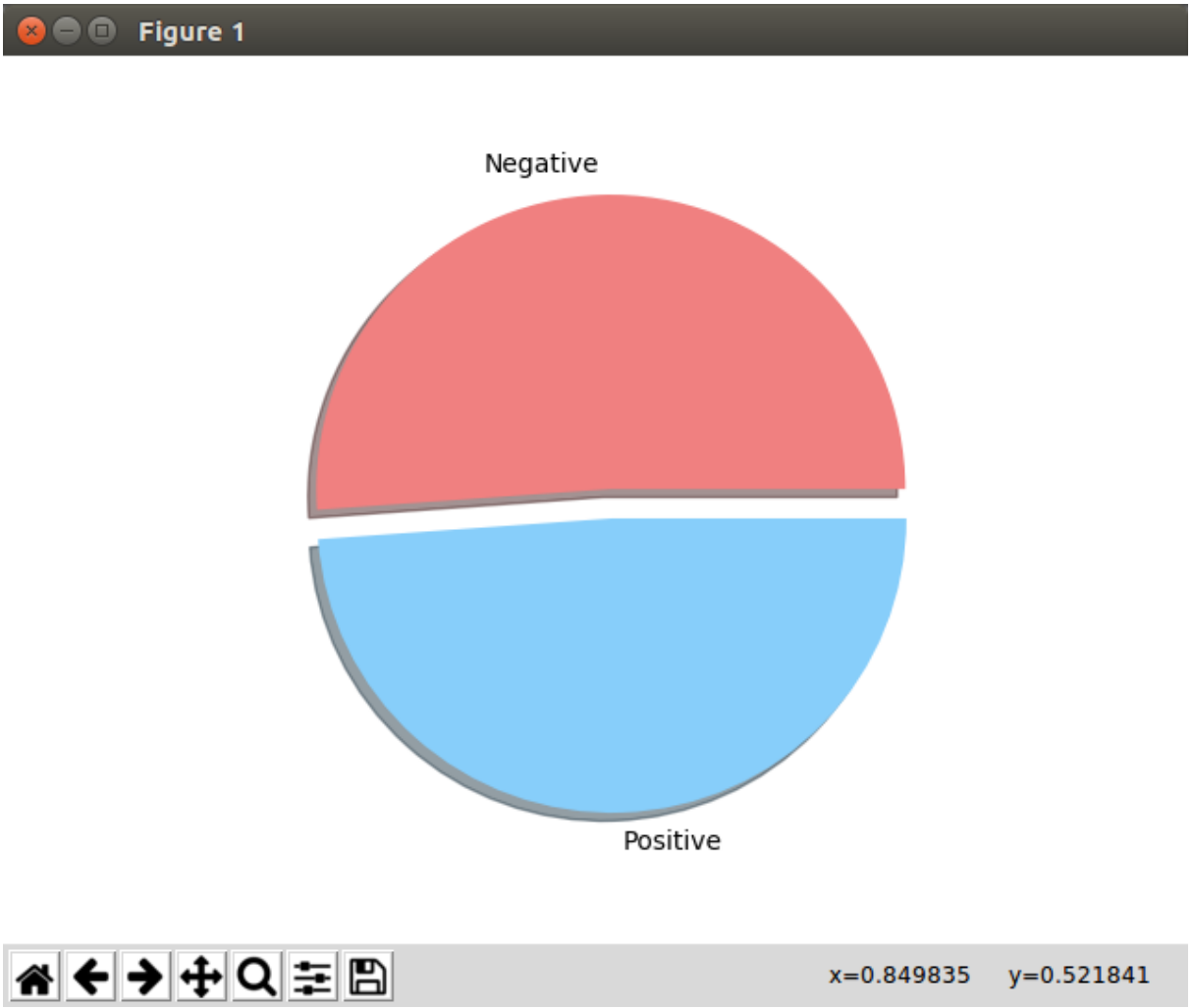

FIGURE 5.5: Plotting the pie chart from the results

In this picture, from the results recieved by the machine, it generates a pie chart for the easy understanding of the user.

Chapter 6

Conclusion and Future Scope

Conclusion

Sentiment analysis or opinion mining is a field of study that analyzes peoples sentiments, attitudes, or emotions towards certain entities. This paper tackles a fundamental problem of sentiment analysis, sentiment polarity categorization. Online product reviews from Amazon.com are selected as data used for this study. A sentiment polarity categorization process has been proposed along with detailed descriptions of each step. Experiments for both sentence-level categorization and review-level categorization have been performed.

Future Scope

- The project can be made more interactive by developing a beautiful User Interface in Django.
- In future, the project can be tailored to satisfy the needs of a particular client by adding the required features or removing the unwanted features.
- The project can also be upgraded to a customized algorithm defined by the developer.
- The project has a brighter scope and we hope, it will also keep satisfying more and more customers in the future.

Chapter 7

Refrences

1. Liu B (2010) Sentiment analysis and subjectivity In: Handbook of Natural Language Processing, Second Edition.. Taylor and Francis Group, Boca.[accesed on 24/02/2017].

2. https://www.tutorialspoint.com/python2.7/ [accesed on 10/03/2017].

3. Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr2(1-2): 1135. [accesed daily].

4. Whitelaw C, Garg N, Argamon S (2005) Using appraisal groups for sentiment analysis [accesed on
20/03/2017].

5. http://textminingonline.com/dive-into-nltk-part-iii-part-of-speech-tagging-and-pos-tagger [accesed on 27/03/2017]

6. http://textminingonline.com/dive-into-nltk-part-x-play-with-word2vec-models-based-on-nltk-corpus [accesed on 04/04/2017]