**Sentiment Analysis**

**On Amazon’s Reviews**

A project report submitted for pre final year of

**Bachelor of Technology**

in

**Computer Science and Engineering**

By

B.Prasad (N130888)

M.Vinod Kumar (N130791)

*Under the Supervision of*

**Mr. Amit Patel**



**Department of Computer Science and Engineering**

**Rajiv Gandhi University of Knowledge Technologies – Nuzvid**

**Nuzvid, Krishna – 521202**

April, 2017

**--------------------------------------------------------------------------------------------------------------------------------------**

**Rajiv Gandhi University of Knowledge Technologies**

**(A.P. Government Act 18 of 2008)**

**RGUKT-NUZVID, Krishna Dist - 521202**

Tele Fax : 08656 – 235557/235150

## CERTIFICATE OF COMPLETION

This is to certify that the work entitled, “Sentiment Analysis on Amazon’s reviews**”** is the bonafied work of ***B.Prasad , ID No: N130888,M.Vinod Kumar , ID No: N130791*** carried out under my guidance and supervision for pre final year project of **Bachelor of Technology** in the department of Computer Science and Engineering under RGUKT IIIT Nuzvid. This work is done during the academic session August 2017 – December 2017, under our guidance.

­­

---------------------------- ---------------------------

**Mr.Amit Patel Mr.Upendra Kumar**

Project SupervisorHead of Department

Lecturer Dept. of CSE Assistant Professor Dept. of CSE

RGUKT IIIT Nuzvid, Nuzvid RGUKT IIIT Nuzvid, Nuzvid

**--------------------------------------------------------------------------------------------------------------------------------------**

**Rajiv Gandhi University of Knowledge Technologies**

**(A.P. Government Act 18 of 2008)**

**RGUKT-NUZVID, Krishna Dist - 521202**

Tele Fax : 08656 – 235557/235150

## CERTIFICATE OF EXAMINATION

This is to certify that the work entitled, “**Sentiment Analysis Using Amazon reviews”** is the bonafide work of ***B.Prasad ID No: N130888,M.Vinod Kumar, ID No: N130791*** and here by accord our approval of it as a study carried out and presented in a manner required for its acceptance in pre final year of **Bachelor of Technology** for which it has been submitted. This approval does not necessarily endorse or accept every statement made, opinion expressed or conclusion drawn, as a recorded in this thesis. It only signifies the acceptance of this thesis for the purpose for which it has been submitted.

­­

---------------------------- ---------------------------

**Mr.Amit Patel Examiner**

Project SupervisorProject Examiner

Lecturer Dept. of CSE Lecturer Dept. of CSE

RGUKT IIIT Nuzvid, Nuzvid RGUKT IIIT Nuzvid, Nuzvid

**--------------------------------------------------------------------------------------------------------------------------------------**

**Rajiv Gandhi University of Knowledge Technologies**

**(A.P. Government Act 18 of 2008)**

**RGUKT-NUZVID, Krishna Dist - 521202**

Tele Fax : 08656 – 235557/235150

## DECLARATION

We, ***B.Prasad, ID No: N130888 M.Vinod kumar, ID No: N130791*** hereby declare that the project report entitle “**Sentiment Analysis using Amazon reviews”** done by us under the guidance of **Mr. Amit Patel M.Tech** is submitted for pre final year of **Bachelor of Technology** in **Computer Science and Engineering** the academic session August 2017 – December 2017 at RGUKT – Nuzvid.

We also declare that this project is a result of our own effort and has not been copied or imitated from any source. Citations from any websites are mentioned in the references.

The results embodied in this project report have not been submitted to any other university or institute for the award of any degree or diploma.

­­

**Date : \_\_\_\_\_\_\_\_\_\_\_\_ B.Prasad [N130888]**

**Place :\_\_\_\_\_\_\_\_\_ M.Vinod Kumar[N130791]**

**ACKNOWLEDGEMENT**

We would like to express our profound gratitude and deep regards to our guide ***Mr. Amit Patel*** for his exemplary guidance, monitoring and constant encouragement throughout the course of this thesis.

We are extremely grateful for the confidence bestowed in us and entrusting our project entitled “Sentiment Analysis on Amazon’s reviews”.

At this juncture we feel deeply honored in expressing our sincere thanks to him for making the resources available at right time and providing valuable insights leading to the successful completion of our project.

We would like to thank RGUKT Nuzvid Director, faculty and staff for their valuable suggestions and discussions.

Last but not least I thank almighty, and I place a deep sense of gratitude to my family members and my friends who have been constant source of information during the preparation of this project work.

B.Prasad,N130888

M.Vinod kumar,N130791

**ABSTRACT**

Users of the online shopping site, Amazon are encouraged to post reviews of the products that they purchase. The reviews provide accessible and plentiful data for relatively easy analysis for a range of applications. Recommendation systems are typically used by companies, especially e-commerce companies like Amazon to help users discover items they might not have found by themselves and promote sales to potential customers. This is a highly-targeted approach which can generate high conversion rate and make it very effective and smooth to do advertisements.

Sentiment analysis refers to the use of [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing), [text analysis](https://en.wikipedia.org/wiki/Text_analytics) and [computational linguistics](https://en.wikipedia.org/wiki/Computational_linguistics) to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to [voice of the customer](https://en.wikipedia.org/wiki/Voice_of_the_customer) materials such as reviews and survey responses, online and social media, and healthcare materials.

This project seeks to apply and extend the current work in the field of natural language processing and sentiment analysis to data retrieved from Amazon. Stanford coreNLP , Sentiword net and SVM Classifier are used to classify a given review as positive or negative. The number of stars a user gives a product is used as labels to perform supervised machine learning. A corpus contains 20,000 product review serves as the dataset of project. Useful features of them that aid in accurate classification are compared to those most useful in classification of remaining product reviews. The features, such as unigrams and bigrams, are compared to one another in their effectiveness in correctly tagging reviews.

**CONTENTS**

1. **Introduction** ……………………………………………... 01
   1. Importance of Sentiment Analysis …………….. 01
2. **Levels of Sentiment Analysis** ..…………………………… 02
3. **Literature Review** ……………………………………….. 04
4. **Our Implementation** …………………………………. 06
   1. Data Gathering ……………………... 06
   2. Splitting Data ................................... 06
   3. Data Pre-processing ………………........... 07
   4. Tagging Data ……………………. 08
   5. Token Scoring ……………………... 09
   6. Feature Extraction ................................... 09
   7. Support Vector Machine …………………... 10
5. **Experimental Results** ……………….. 12
6. **Conclusion and Future Scope** ……………………………. 13

**References** ……………………………………………............ 14

**LIST OF FIGURES**

4.1 Sample data …………………………………… 14

4.2 Splitting of data …………………………………… 15

4.3 Data pre-processing …………………………………… 16

4.4 Data Tagging …………………………………… 17

4.5 Data Scoring …………………………………… 18

4.6 Feature Extraction ………….……………………….. . 19

4.7 Simple SVM …………………………………… 20

**ABBREVIATIONS**

JSON Java Script Object Notation

SA Sentiment Analysis

Kernel Function used to Convert m-d to n-d

NLP Natural Language Processing

RBF Radial Basis Function

SVM Support Vector Machine

POS Parts Of Speech

ADV Adverb

V Verb

ADJ Adjective

**CHAPTER 1**

Introduction

**Sentiment analysis** (known as **opinion mining** or **emotion AI**) refers to the use of [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing), [text analysis](https://en.wikipedia.org/wiki/Text_analytics), [computational linguistics](https://en.wikipedia.org/wiki/Computational_linguistics), and [biometrics](https://en.wikipedia.org/wiki/Biometrics) to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to [voice of the customer](https://en.wikipedia.org/wiki/Voice_of_the_customer) materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from [marketing](https://en.wikipedia.org/wiki/Marketing) to [customer service](https://en.wikipedia.org/wiki/Customer_relationship_management) to clinical medicine..

* 1. **Importance of Sentiment Analysis**

“What other people think” has always been an important piece of information for most of us during the decision-making process. Long before awareness of the World Wide Web became widespread, many of us asked our friends to recommend before we buy any thing. But the Internet and the Web have now made it possible to find out about the opinions and experiences of those in the vast pool of people .Conversely, more and more people are making their opinions via the Internet.

We hasten to point out that consumption of goods and services is not the only motivation ,a need for political information is another important factor. For example, in a survey of over 2500 American adults, the 31% of Americans — over 60 million people — that were 2006 campaign internet users, defined as those who gathered information about the 2006 elections online and exchanged views via email.

Marketers have always needed to monitor media for information related to their brands — whether it’s for public relations activities, fraud violations, or competitive intelligence. But fragmenting media and changing consumer behaviour have crippled traditional monitoring methods. Surveys estimate that 75,000 new blogs are created daily, along with 1.2 million new posts each day, many discussing consumer opinions on products and services. Tactics [of the traditional sort] such as clipping services, field agents, and ad hoc research simply can’t keep pace.

Sentiment Analysis plays an important role in doing all of the Above thing. We can know our brand in the social media by collecting and appliying SA on data.

**CHAPTER 2**

Levels Of Sentiment Analysis

Sentiment Analysis can be done at few Levels based on the user need and their data. Each Level will be useful in some application. They vary from size, complexity , inconsistencies, redundancy of data with which they are dealing with. Having a Sentiment analysis at one level and to move another level involved in increment of complexity as well as time consuming.

Levels of Sentiment Analysis…..

1. Phrase Level
2. Sentence Level
3. Para Level
4. Document Level
5. Aspect Level
6. Objectivity Level

### Phrase Level

This is the simplest form of sentiment Analysis in which we will take an phrase and classify whether it belongs to +ve, -ve or neutral. It has no much complexity involved and can be done be by simple algorithms like dictionary methods.

### Sentence Level

This level is confined to sentence . The Sentence is the collection and words or phrases. We will apply Sentiment Analysis to the Sentence at once . Here we should take about all the phrases that are involved in the sentence not a single word or phrase.

### Paragraph Level

Paragraph is the collection of sentence ,which in turn formed with collection of phrases. This is level gives us a challenge to handle with more than one sentence at an instant. This would be the complex task than above one and needs to identify different features that are involved in each sentence. Splitting Para into Multiple sentences and apply sentence is Level SA is allowed but brings down performance to ground and moreover it is time consuming. Our project is dealing with paragraph level SA . We’ll treat each review as a paragraph and apply SA.

### Document Level

Document will consists of several paragraphs. Finding out whole document level summary is the challenging work. Sentiment Analysis on Document will classify the whole document into one of classes +ve,-ve and neutral.

### Feature/Aspect Level

It refers to determining the opinions or sentiments expressed on different features or aspects of entities, e.g., a digital camera, or a bank. A feature or aspect is an attribute or component of an entity, e.g., the screen of a cell phone, the service for a restaurant. The advantage of feature-based sentiment analysis is the possibility to capture nuances about objects of interest. Different features can generate different sentiment responses, for example a hotel can have a convenient location, but mediocre food. This problem involves several sub-problems, e.g., identifying relevant entities, extracting their features/aspects, and determining whether an opinion expressed on each feature/aspect is positive, negative or neutral.

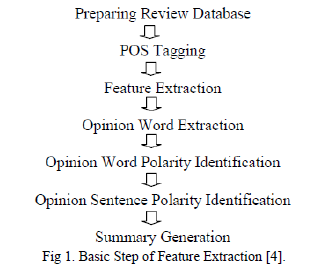
### Subjectivity/Objectivity Level

This task is commonly defined as classifying a given into one of two classes: objective or subjective. This problem can sometimes be more difficult than polarity classification. The subjectivity of words and phrases may depend on their context and an objective document may contain subjective sentences (e.g., a news article quoting people's opinions). Moreover, results are largely dependent on the definition of subjectivity used when annotating texts. However , removing objective sentences from a document before classifying its polarity helped improve performance.

**CHAPTER 3**

Literature Review

All Information in the world can be broadly classified into mainly two categories, facts and opinions. Facts are objective statements about entities and worldly events. On the other hand opinions are subjective statements that reflect people’s sentiments or perceptions about the entities and events . Maximum amount of existing research on text and information processing is focused on mining and getting the factual information from the text or information. Before we had WWW we were lacking a collection of opinion data, in an individual needs to make a decision, he/she typically asks for opinions from friends and families. When an organization needs to find opinions of the general public about its products and services, it conducted surveys and focused groups. But after the growth of Web, especially with the drastic growth of the user generated content on the Web, the world has changed and so has the methods of gaining ones opinion. One can post reviews of products at merchant sites and express views on almost anything in Internet forums, discussion groups, and blogs, which are collectively called the user generated content . As the technology of connectivity grew so as the ways of interpreting and processing of users opinion information has changed. Some of the machine learning techniques like Naïve Bayes, Maximum Entropy and Support Vector Machines has been discussed in the paper [1]. Extracting features from user opinion information is an emerging task.



In Fig 1,a generic model of feature extraction from opinion information is shown, firstly the information database is created, next POS tagging is done on the review, next the features are extracted using grammar rules such as adjective + noun or so on, as nouns are features and adjectives are sentiment words. Next Opinion words are extracted followed by its polarity identification . Some models also calculate sentence polarity for accuracy. Lastly the results are combined to obtain a summary. Many algorithms can be used in opinion mining such as Naïve Bayes Classification, Probabilistic Machine Learning approach to classify the reviews as positive or negative, have been used to get the sentiment of opinions of different domains such as movie , Amazon reviews of products In our work we have used reviews of amazon extracted from private website. We studied all the reviews and got to know that there are many reviews in which the user talks about the service provided by amazon and its sellers. So we decided to classify reviews into service, product and feature based reviews. We also found that the sentiment of each review is very obvious, the review rating provided by the user mirrors what the user writes as his/her review, i.e. if the user writes something bad definitely the overall rating the user gives is either 1 or 2 out of 5. This is from our study of a set of amazon reviews on various products. Our work mainly concentrates on feature extraction and finding out the sentiment of the particular feature. We have used POS tagging technique on sentence level. In our approach we have made certain rules using the tags of particular word and using list of words with respective sentiment value to find the feature and then getting the appropriate sentiment from it. The Sentiment model that we have proposed is designed based on the features of the amazon reviews.

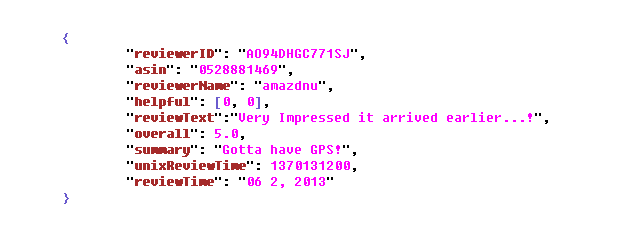
**CHAPTER 4**

Our Implementation

In this chapter you will be explained briefly about our proposed methodology and their practical implementation. In this section we’ll going to present each phase that is involved in the project ,given with pictures and packages that we have used throughout the whole project. The whole Project is done within the Java, object-orinted programming .

* 1. **Data Gathering**

We Gathered Data related to the Amazon’s review Data set from the corpus which resides in the given link http://jmcauley.ucsd.edu/data/amazon/ . The Data which was downloaded is in the format of JSON. Each review is reside in the downloaded document is within the java script object format.

 Fig 4.1 Sample Data

In Fig 4.11 is example of review, which consists of all details of a review like time, username, review text ,summary and rating etc..,

* 1. **Splitting Data**

As we want to make an classifier which will be trained well by making some model , it need some part of data as a training set. The classifier, as a part of machine learning will make a model so that it could predict the labels of testing data. Thus, We need two sets of data called Testing data and Training Data.

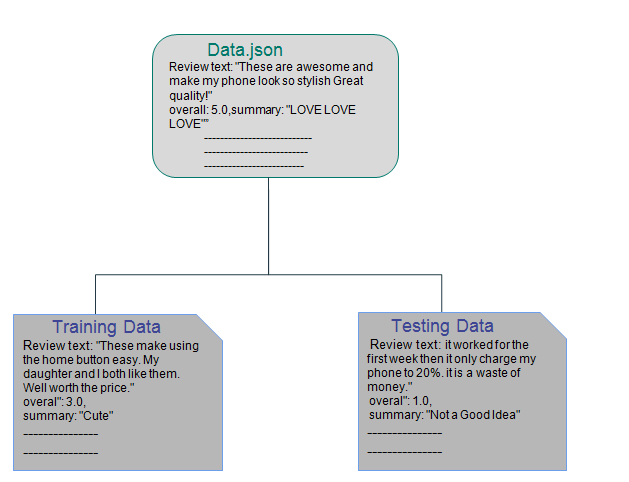


Fig 4.2 Splitting of Data

In This Project we use Support Vector Machine as a classifier. Below is the figure 4.2 which depicts the whole procedure .

* 1. **Data Pre-Processing**

In Fig 4.1 , we have seen that a data gathered have many contents. But of all we need only few entries which are useful for our project. The Data pre-processing part will make new documents , contains the contents only we required. For our Project we need Review Text , Summary and Overall rating such as 3,4,5….

In order to parse data as it was in JSON format , we need the library called simple\_json.jar . It will help us to get the data in format of java script objects in java programming and then we can access each content with it’s key since JSON will have contents in the form of key, value pairs .

With the help of simple\_json parser , Scan the whole input document and extract the review Text , Summary and overall rating which we need.

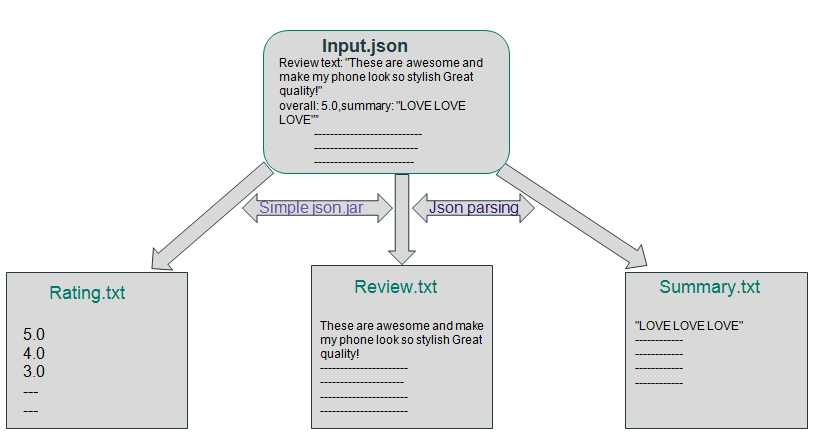
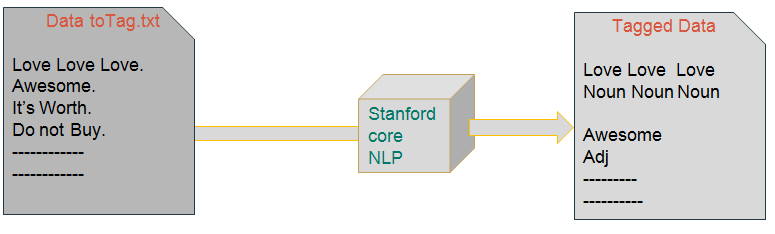


Fig 4.3 data pre-processing

In Fig 4.3 , we have ratings of each review.SVM is the supervised learning , which uses the labeled data. In our Project , we treat the review’s with rating as 1 and 2 are belongs to +ve class whereas review’s with 4 and 5 are treated as –ve class since the user’s rating will be biased on their sentiment only. Labeling the input (test or train) data is dwell in this phase only. Labels of Train data used in training process of SVM whereas we’ll hide Labels of Test data ,which are used to compare later after their prediction with classifier.

* 1. **Tagging Data**

From pre-processing we got data that was labeled with each review having it’s review text and summary. SVM does not deal with raw text lest it works on the feature vector. The Working of SVM explained briefly in Sec 4.10. In order to know the influence of a word we has to first know the POS for which it belongs to . For that purpose we are going to use Stanford Core NLP library which was developed by the people of Stanford as like NLTK Package in python.

 Fig 4.4 Data Tagging

Core NLP has features like split, lemmatization , named entity recognition and POS tagging etc.., We have used sentence split for splitting of reviews into sentences and POS tagger for tagging the each word. After tagging data i.e., both review and summary data separately , we get tokens with their POS tags as shown in Fig 4.4.

* 1. **Token Scoring**

To know each token’s influence in sentence we can give the scoring to the tokens. The score would be based on the subjectivity , objectivity , probability of +ve and –ve. SentiWord Net is such a corpus at which we will find the scores to almost every word which is available online at following link <http://sentiwordnet.isti.cnr.it/> . It is formed with plenty of words in the form of network like structure and have having each word with its equivalents in that network. They are called as synsets.

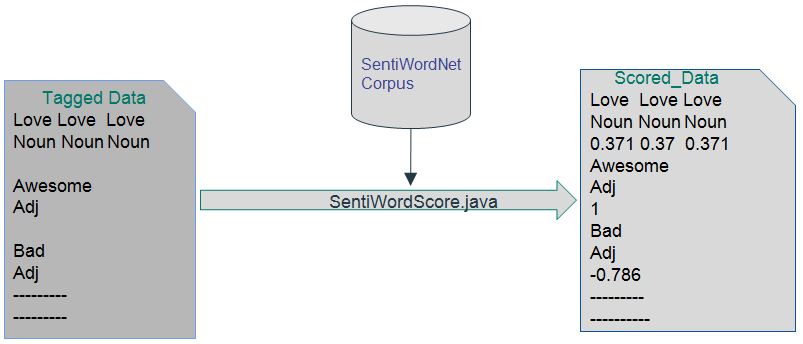
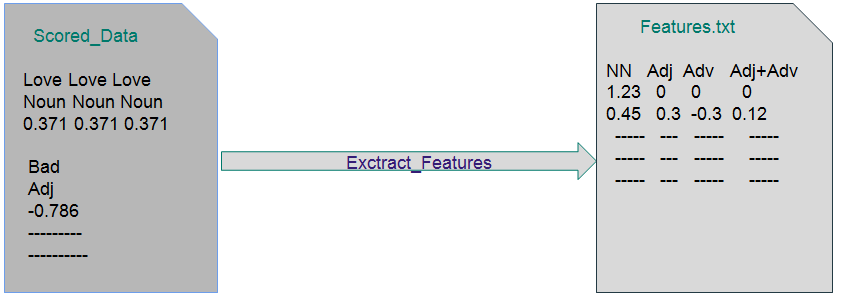


Fig 4.5 Data Scoring

Although we deriving all the words with scores. It turns out that we does not need all of them as they will not make opinion in the given sentence. It is not exactly we would do all the phases separately in sequential manner . Some phases are combined with others to increase parallelism and to reduce time complexity. The Layered Structure which we are explained here can be viewed as Proposes Method but coming to the Practical Implementation some phases are tied together as said earlier.

* 1. **Feature Extraction**

In sec 4.5 you have been explained about scoring and being said that all words are not influential. We would decide what are the features that are providing opinions in this phase. Using unigrams, Bigrams and tri-grams we can train our machine to classify the data. But In our Research we found out that trigrams are not influencing that much extent since they were implicitly show their influence in uni and bi –grams . So, In our project we are extending upto bigrams only. The Adjective is the best influencing part of all the sentence , so we would take care about the adj and adj+noun.

 Fig 4.6 Feature Extraction

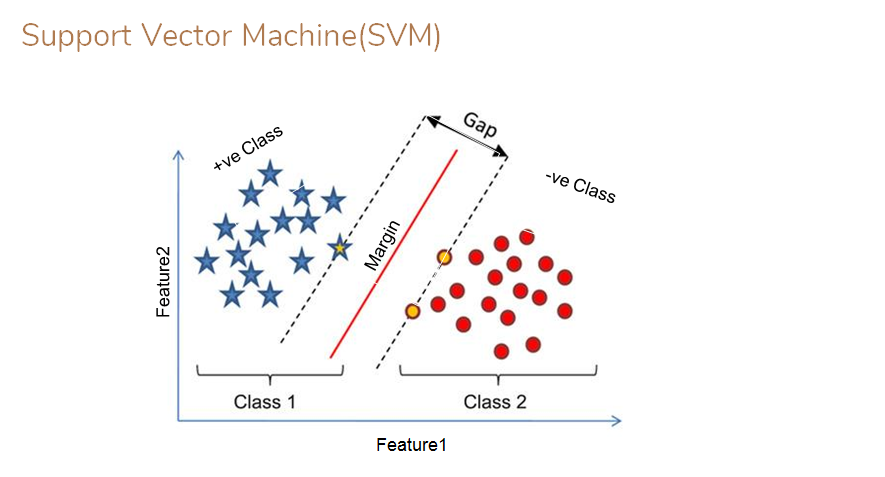
In Fig 4.6 we can see the kinds of features we have taken. It has to be noticed that we need to perform feature extraction on both Review Text and Summary. Later when we train the system , we could load both separate feature vectors as one. At the same time , we should also for the Test data as well through all above phases.

* 1. **Support Vector Machine(SVM)**

In [machine learning](https://en.wikipedia.org/wiki/Machine_learning), **support vector machines** (**SVMs**, also **support vector networks**) are [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) models with associated learning [algorithms](https://en.wikipedia.org/wiki/Algorithm) that analyze data used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis). Given a set of training examples which are labeled , an SVM a model that predicts test data to one category or the other, making it a non-[probabilistic](https://en.wikipedia.org/wiki/Probabilistic_classification) [binary](https://en.wikipedia.org/wiki/Binary_classifier) [linear classifier](https://en.wikipedia.org/wiki/Linear_classifier) . An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible and is called **HYPERPLANE**. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the [kernel trick](https://en.wikipedia.org/wiki/Kernel_trick), implicitly mapping their inputs into high-dimensional feature spaces.

When data are not labeled , supervised learning is not possible, and an [unsupervised learning](https://en.wikipedia.org/wiki/Unsupervised_learning) approach is required, which attempts to find natural [clustering of the data](https://en.wikipedia.org/wiki/Data_clustering) to groups, and then map new data to these formed groups. The clustering algorithm which provides an improvement to the support vector machines is called **support vector clustering** and is oftenused in industrial applications either when data are not labeled or when only some data are labeled as a preprocessing for a classification pass.

In Fig 4.8 It has taken that a simple two feature train data and that is to linearly separable. From the diagram it is clear that SVM needs only two features that are at edges of each plane to construct a hyper-plane. The features that are support for constructing hyperplane are called support vectors. The hyperplane(or line in 2-D) is equi-distant from the two planes so that it  Fig 4.7 Simple SVM

could maximize the prediction. If Features are non-linearly separable we use kernel trick. In our Project we are taking eight features to train machine and using RBF kernel to convert data into linearly separable.

SVM is available as API’s in java and as libraries in Python like sklearn etc.., We are using Stanford JNI\_SVM light interface to train and test our Features.

**CHAPTER 5**

Experimental Results

The size of data initially we have taken is consists of 200k reviews. Due to Memory and System configurations we have reduced it to the 20k review. Of which we split the data into train data and test data in the ratio of 90:10 respectively. We got 2000 reviews as a Test Data set and remaining treated as the Train Set.

The time taking for tagging and scoring 20k reviews is around 4+hours. By the use of **Divide and Conquer** and files as buffers we are able to reduce the time to 15 minutes. The simultaneous computation of Stanford Core NLP and SentiWord Net remains in lack of sufficient memory.

We are using Stanford JNI SVM light API for SVM classifier. Training the SVM and testing our test set results in 75.9% accuracy.

**CHAPTER 6**

Conclusion And Future Scope

**6.1 Conclusion**

This report presents the implementation of proposed method of Sentiment Analysis. We have implemented Sentiment Analysis practically with help of JAVA, Stanford core NLP, Json parser ,Senti word Net and Stanford SVM light.

We have reduced the time complexity by using principle of **Divide and Conquer** and using files as synchronizing buffers. We store Intermediate results in txt formats so that system’s space complexity will be reduced.

**6.2 Future Scope**

In this Mini- Project we are able to perform review level classification with help of SVM . In future we will going calculate accuracy measures like precision and recall etc., and perform k-folding cross validation to increase accuracy of training model.

Simple Review may contains several noisy data and the content may not be related to the Aspect which they are reviewing for . Now , Aspect Level will come into role which gives us an Challenging task to do and trickier than earlier.

References

[1 ] ChandraKala1 and C. Sindhu2, “OPINION MINING AND SENTIMENT CLASSIFICATION: A SURVEY,” .Vol .3(1),Oct- 2012,420-427.

[2]*Stanford core NLP API*. (2017, August 23). Retrieved  from h[ttps://stanfordnlp.github.io/CoreNLP/api.html](https://stanfordnlp.github.io/CoreNLP/api.html)

[3] *Senti Word Net Demo*. (2017, Sept 17). Retrieved  from <http://sentiwordnet.isti.cnr.it/code/SentiWordNetDemoCode.java>

[4]. Callen Rain , "Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning" crain1@swarthmore.edu.

[5]. Jianfeng Hu , Bo Zhang ” Product Recommendation System” Stanford Project Review(2012 .Dec)

[6]. Bing Liu Science of Illinois at Chicago “ Sentiment Analysis and Opinion Mining” AAAI-2011 Tutorial (2012 ,may) 0-100