

# A study on Sentiment Analysis

Himanshu Kumar Meher



Department of Computer Science and Engineering  
National Institute of Technology Rourkela  
Rourkela-769 008, Odisha, India.

**A study on Sentiment Analysis**

*Thesis submitted in partial fulfilment  
of the requirements for the degree of*

**Bachelor of Technology**

*in*

**Computer Science and Engineering**

*by*

**Himanshu Kumar Meher**

**(Roll: 111CS0136)**

*under the guidance of*

**Prof. Dr.Korra Satya Babu**



**Department of Computer Science and Engineering  
National Institute of Technology Rourkela  
Rourkela-769 008, Odisha, India.**

**March' 2015**



Department of Computer Science and Engineering  
**National Institute of Technology Rourkela**  
Rourkela-769 008, Orissa, India.

May 9, 2015

## DECLARATION

This thesis is a presentation of my original research work. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgment of collaborative research and discussions. I hereby declare that this thesis is my own work and effort and that it has not been submitted anywhere for any award. The interpretations put forth are based on my reading and understanding of the original texts and they are not published anywhere in the form of books, monographs or articles. The other books, articles and websites, which I have made use of are acknowledged at the respective place in the text. For the present thesis, which I am submitting to NIT Rourkela, no degree or diploma or distinction has been conferred on me before, either in this or in any other University. I bear all responsibility and prosecution for any of the unfair means adopted by me in submitting this thesis.

**Date**

**Signautre**

# Acknowledgment

I owe deep gratitude to the ones who have contributed greatly in completion of this thesis.

Foremost, I would also like to express my gratitude towards my project advisor, Prof. Korra Satya Babu, whose mentor-ship has been paramount, not only in carrying out the research for this thesis, but also in developing long-term goals for my career. His guidance has been unique and delightful. He provided his able guidance whenever I needed it. I would also like to thank my Ph.D. mentor Mr.Santosh Bharati who helped me greatly by being a source of knowledge to me.

I would also like to extend special thanks to my project review panel for their time and attention to detail. The constructive feedback received has been keenly instrumental in improvising my work further.

My parents receive my deepest love for being the strength in me.

*Himanshu Kumar Meher*

Roll No. 111CS0136

# Abstract

Now days the growth of social websites, blogging services and electronic media contributes huge amount of user give messages such as customer reviews, comments and opinions . Sentiment Analysis is important term of referred to collection information in a source by using NLP, computational linguistics and text analysis and to make decision by subjective information extracting and analyzing opinion, identifying positive and negative reviews measuring how positively and negatively an entity( public ,organization, product) is involved. Sentiment analysis is the area of study to analyze peoples reviews, emotion, attitudes and emotion from written languages. We concentrate on field of different opinion classification techniques, performed on any data set. Now days most popular approaches are Bag of words and feature extraction used by researchers to deal with sentiment analysis is used by politician, news groups, manufactures organization, movies, products etc.

**Keywords:**[Text mining, Sentiment analysis, Natural language processing].

# Contents

<b>Certificate</b>	<b>ii</b>
<b>Acknowledgement</b>	<b>iii</b>
<b>Abstract</b>	<b>iv</b>
<b>List of Figures</b>	<b>vi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Sentiment analysis: . . . . .	2
1.2 Opinion mining: . . . . .	2
1.3 Why opinion mining?: . . . . .	2
<b>2 Literature Review</b>	<b>3</b>
2.1 Opinion mining technology . . . . .	3
2.2 Steps in opinion mining approach: . . . . .	4
<b>3 Proposed Work</b>	<b>7</b>
3.1 EWGA algorithm . . . . .	8
<b>4 Results</b>	<b>11</b>
4.1 dataset . . . . .	13
<b>5 Conclusion</b>	<b>17</b>
<b>Bibliography</b>	<b>18</b>

# List of Figures

2.1	system design	6
3.1	EWGA Illustration	9
4.1	Collection of Hotel dataset	11
4.2	Extraction of n-gram	12
4.3	select data	12
4.4	POS	13
4.5	sentiment	14
4.6		15
4.7		16

# Chapter 1

## Introduction

Another aspect of web is e-commerce which has been on the rising since the 21st century started. More items are being sold nowadays on the web and the customer database is likewise expanding. The online shippers ask their customers to give their valuable feedback when they purchase some item in order to upgrade consumer loyalty and shopping background. With more people purchasing items through web the number of surveys for each item is growing rapidly leading to huge amount of data to be processed. The item makers may face difficulty because of the amount of inputs coming through from the users. Besides, customers sometimes may get a negative feedback just by going through one or two feedbacks. [1]

In this section we discuss about Sentiment analysis and opinion mining ,how to analyze In simple terms, opinion mining involves making a system to use of reviews posted by user so as to improve upon the products features. Given a set of reviews our task involves the following: Identify features of the product on which customers have expressed their opinion (called product features). We use techniques like data mining and natural language processing in order to mine the features. [2]

For each feature, partition the entire set of reviews into positive or negative reviews. To decide what the opinion orientation is we perform the three subtask: I. Identify a set of adjectives normally used to express opinions using natural language processing method. These are called opinion words.

II. For each opinion word we determine its semantic orientation.

III. Decide the opinion orientation for each sentence.

Generating a summary out of the discovered information.



## **1.1 Sentiment analysis:**

Sentiment classification is a technique to focus on the sentiments or opinions expressed in an article or conveyed orally. The term sentiment includes emotions, conclusions, behaviour and others. In this report, we concentrate on human readable text writing on the e-commerce sites.

## **1.2 Opinion mining:**

Opinion mining involves analysing opinions, sentiments or mentality of the writer from the written text. Opinion mining uses the concepts of NLP, data mining and machine learning to perform this task. This section involves analysing requirement for opinion mining. In the next segments, we concentrate on sentiment mining assignments and present a review.

## **1.3 Why opinion mining?:**

Online opinions have indirect influence on the business of several e-commerce sites. Those sites market their products and the web users go through the reviews of the product before buying that product. Many organizations utilize opinion mining systems to track customer reviews of products sold online.

Opinion mining is an incredible way of maintaining focus on several business trends related to deals administration, status management and also advertising. Pattern prediction is also done using the opinion of the customers.

# Chapter 2

## Literature Review

In this section a brief introduction about the previous work done in this field is given. A brief summary about the types of sentiment analysis and methods of sentiment classification is also given.

### 2.1 Opinion mining technology

The various terms used in opinion mining are given below:

**Fact:** A certainty is that which has genuinely happened or is truly the case.

**Opinion :** A feeling is a perspective or judgment framed about something, not so much in light of truth or information.

**Subjective Sentence:** A sentence or a content is subjective or stubborn on the off chance that it really demonstrate ones emotions. [3]

**Target Sentence:** A target sentence demonstrates a few actualities and known data about the world.

**Thing:** an individual article or unit, particularly one that is a piece of a rundown, gathering, or set.

**Survey:** An audit is a content containing an arrangement of words that has sentiments of client for a particular thing. A survey may be subjective or objective or both.

**Known Aspects:** Known angles are default perspectives gave by the certain site for which clients independently give appraisals.

**Conclusion:** Sentiment is an extremity term that infers to the heading in which an

idea or feeling is communicated. We utilize assessment in a more particular sense as a conclusion about an angle. For instance, astounding is an opinion for the characteristic 'battery life' in the sentence

"This portable has great battery life".

**Conclusion Phrase:** An assessment expression is a couple of head term and modifier. Generally the head term is a competitor angle, and the modifier is an assumption that communicates some conclusion towards this viewpoint.

**Sentiment Polarity:** Opinion polarity or Subjectivity Orientation means the extremity communicated by the client or client regarding numerical qualities.

**Polarity:** Polarity is a two way introduction scale. In this, a notion can be either positive or negative.

**Rating:** Most of the auditing sites utilization star appraisals for communicating extremity, introduced by stars in the reach from 1 to 5 which are called evaluations.

**General Rating:** All the Internet shopping sites request that clients give a general rating for the item that they as of now purchased saying the general nature of the utilized the Grammatical feature. (PoS) Tag: PoS labeling is extremely helpful in Opinion Mining procedure. When we have to dissect a report or a sentence first we need to concentrate the subjective data from the archive or that specific sentence. PoS labeling helps us in getting subjective words like Nouns, Verbs, Adverbs and Adjectives. In the wake of extricating these words we can perform different activities on these and we can reach a conclusion.

## 2.2 Steps in opinion mining approach:

Pang e al. has grouped the significant issues of opinion mining into three classes : sentiment polarity identification, subjectivity detection, and joint topic-sentiment analysis. One of the most important classifications for opinion mining tasks include: document level, sentence level and phrase level.

There is one more important classification called the feature level opinion mining. Along with that we introduce document level opinion mining and sentence level opinion mining.

Document level opinion mining [4]

Document level tasks primarily deals with classification issues where the available document has to be arranged into a set of predefined clauses. In sentiment analysis, a document is classified as positive, negative or neutral depending upon the polarity of subjective information present in the document. Opinion quality deals with the usefulness of the opinion and whether the opinion is a spam or not.

Subjectivity analysis :

This kind of analysis decides whether a document makes an opinion or not. Precisely, it classifies the document as objective or subjective. Supervised learning is used for this purpose. Future research concentrates on making this analysis process automated.

Sentiment analysis:

This kind of analysis tries to discover the general sentiment of the user as can be understood from the content. It expects that the document is subjective. It figures out the polarity (positive or negative) of the document. Online reviews of the products are used as training information. Nave Bayes, Maximum Entropy classification and SVM are techniques used for analysis.

Sentence-level opinion mining In this type of mining, sentence are analysed to find out the quality of the product. Words such as good, best are identified and polarity is decided according to that. Opinion mining refers to comparing two products of similar category or sometimes we may compare only certain attributes of the different products.

Feature level opinion mining

Feature level opinion mining is useful when an user or customer wants to know about a particular feature of a product. For example many customers search for mobile phones with high mega pixel camera. For such a scenario feature level opinion mining is appropriate. [3]

Extract object features:

In this step, features of a product are extracted from comments and reviews. Procedures such as extracting nouns and noun phrases are used for this purpose.

Determine polarity of opinions on features:

After the feature extraction process our task is to perform semantic analysis on the resulting features which gives information about the each and every feature which customer liked or disliked. [5]

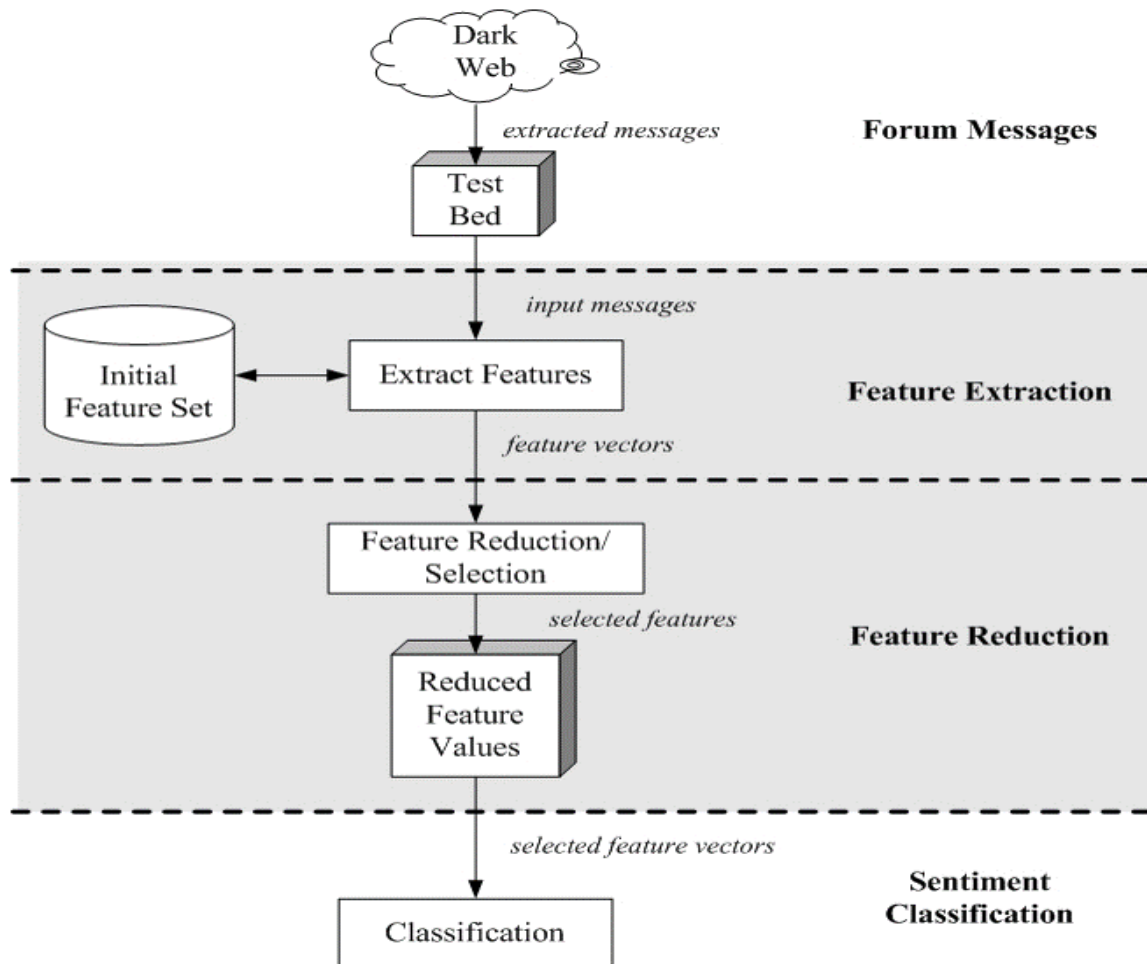


Figure 2.1: system design

# Chapter 3

## Proposed Work

Create word net dictionary: In this type of document, all positive words are written out separately and all negative words are written out at one place.

Extraction of dataset: First dataset of publicly available product reviews were downloaded from the internet and then the passage extraction framework identifies important sections of the text which is most representative of the content of the document. More specifically, this step involves identifying and extracting those specific product features and the opinions on them. Lastly the results are summarized. [6]

Application of n-grams:

Each of the product reviews involves adjectives or adjectives along with other parts of speech. In order to find these information concept of unigrams and bigrams was used. Unigram involves a single word (adjective) which needs to be extracted. For example words like good or bad are unigram which depicts the positive or negative polarity of the review. Bigrams involves two adjacent words like too good or very bad in which the first word if taken will not reveal much information about the sentiment but when we consider the next word then the sentiment of the text can be well understood.

Performing feature selection:

Feature selection is the process of removing redundant features. All the features/words are not applicable for analysis and needs to be removed. For optimization of time and space, the number of features in the feature set has to be decreased. The features of the feature set are filtered using the concept of mutual information.

Mutual information is a quantity that usually measures the mutual dependence of the two random variables. Formally, the mutual information of two discrete random variable  $X$  and  $Y$  can be defined as:  $I(X,Y) = \sum P(x,y) \log P(x,y)/P(x)p(y)$

Find semantic orientation of the document:

Support Vector Machine (SVM) is used to classify the sentiment of the review. A classification technique is a method of building classification models from an input training data set. The model generated should fit the input data correctly and correctly predict the class labels of the test set with as high accuracy as possible. In machine learning, also bolster vector machines (SVMs, likewise bolster vector networks) are regulated learning models with related learning calculations that dissect information and perceive examples, utilized for characterization and relapse investigation. Given an arrangement of preparing cases, every stamped as having a place with one of two classes, a SVM preparing calculation fabricates a model that allocates new illustrations into one class or the other, making it a non-probabilistic parallel straight classifier. A SVM model is a representation of the cases as focuses in space, mapped so that the illustrations of the different classifications are separated by a reasonable crevice that is as wide as could reasonably be expected. New samples are then mapped into that same space and anticipated to fit in with a class in view of which side of the crevice they fall on. The positive sentiments are classified as 1 while the negative sentiments are classified as 0.

### 3.1 EWGA algorithm

1. Information Gain is used to derive feature weights.
2. Include the features selected by IG as part of initial Genetic Algorithm solution population.
3. Based on fitness functions, solutions are evaluated and selected.
4. For each solution pair, find the point at which the IG difference becomes maximum. Perform crossover at that point.
5. Mutate solutions depending upon IG feature weights.
6. Repeat steps 3-5 until stopping criterion is satisfied.

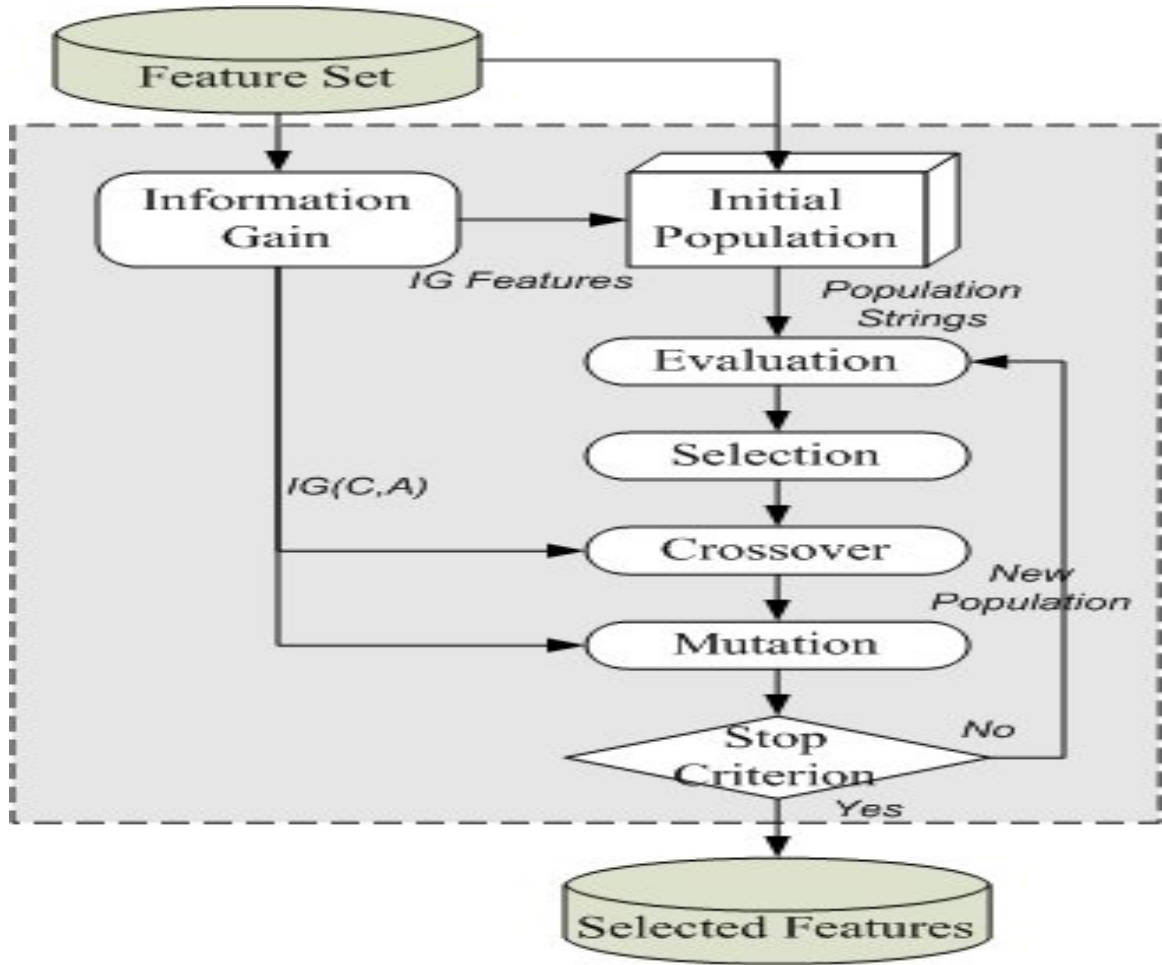


Figure 3.1: EWGA Illustration

Information Gain. For data pick up (IG) we utilized the Shannon entropy measure in which:

$$IG(C,A) = H(C) - H(C-A)$$

where:

$IG(C, A)$  data pick up for highlight A;

$H(C) = -\sum_{i=1}^n p(C=i) \log_2 p(C=i)$  (entropy crosswise over feeling classes C)

$H(C-A) = -\sum_{i=1}^n [p(C=i-A) \log_2 p(C=i-A)]$  (specific highlight contingent entropy) =q

n= complete num of assessment classes.; Entroy(IG)=-plog<sub>2</sub>p-qlog<sub>2</sub>q In the event that the quantity of positive and negative assessment messages is measure up to



, $H(C)=1$ .

IG will differ along the scope of 0 to 1 and all  $IG(C,A) \leq 0.0025$

Initial population: We speak to every arrangement in the population utilizing a paired string of length equivalent to equivalent to the aggregate number of highlight, with every two fold string character speaks to a solitary highlight .1=selected highlight and 0=discarded one. example ,an answer string speaking to five applicant highlight "10011".so 1,4 and 5 are chosen and other are disposed of.

In EWGA , $n-1$  arrangement strings are arbitrarily produced while the IG arrangement highlights are utilized as the last arrangement string in the starting population.

Evaluation and Selection: we utilize grouping precision as the wellness used to assess the nature of every arrangement. Thus, for every genome in the populace, tenfold cross validation with SVM is utilized to get to the wellness of specific arrangement.

Crossover: From the  $n$  arrangement strings in the population (i.e  $n/2$  sets) ,strings sets are arbitrarily chosen for hybrid in view of a hybrid likelihood  $P_c$ .

$S=010010$   $S=010-010$   $S=010100$   $T=110100$   $T=110-100$   $T=110010$   $X=3$   $\arg \max - \sum_{a=1}^x IG(C,A)(S_a - T_a) + \sum_{a=x}^m IG(C,A)(T_a - S_a) -$   
where

$S_a = a$  th character in arrangement string  $S$ ;

$T_a = a$  th character in arrangement string  $T$ ;

$m =$  total no of features;

$x =$  cross over point in arrangement pair  $S$  and  $T, 1 \leq x \leq m$

Mutation:

GA mutation operator randomly mutates individual feature characters in solution string based on a mutation probability  $P_m$ . 0 gives 1 and 1 gives 0.

Evaluation : Accuracy = Number of correctly classified Documents divided by Total number of Documents

# Chapter 4

## Results

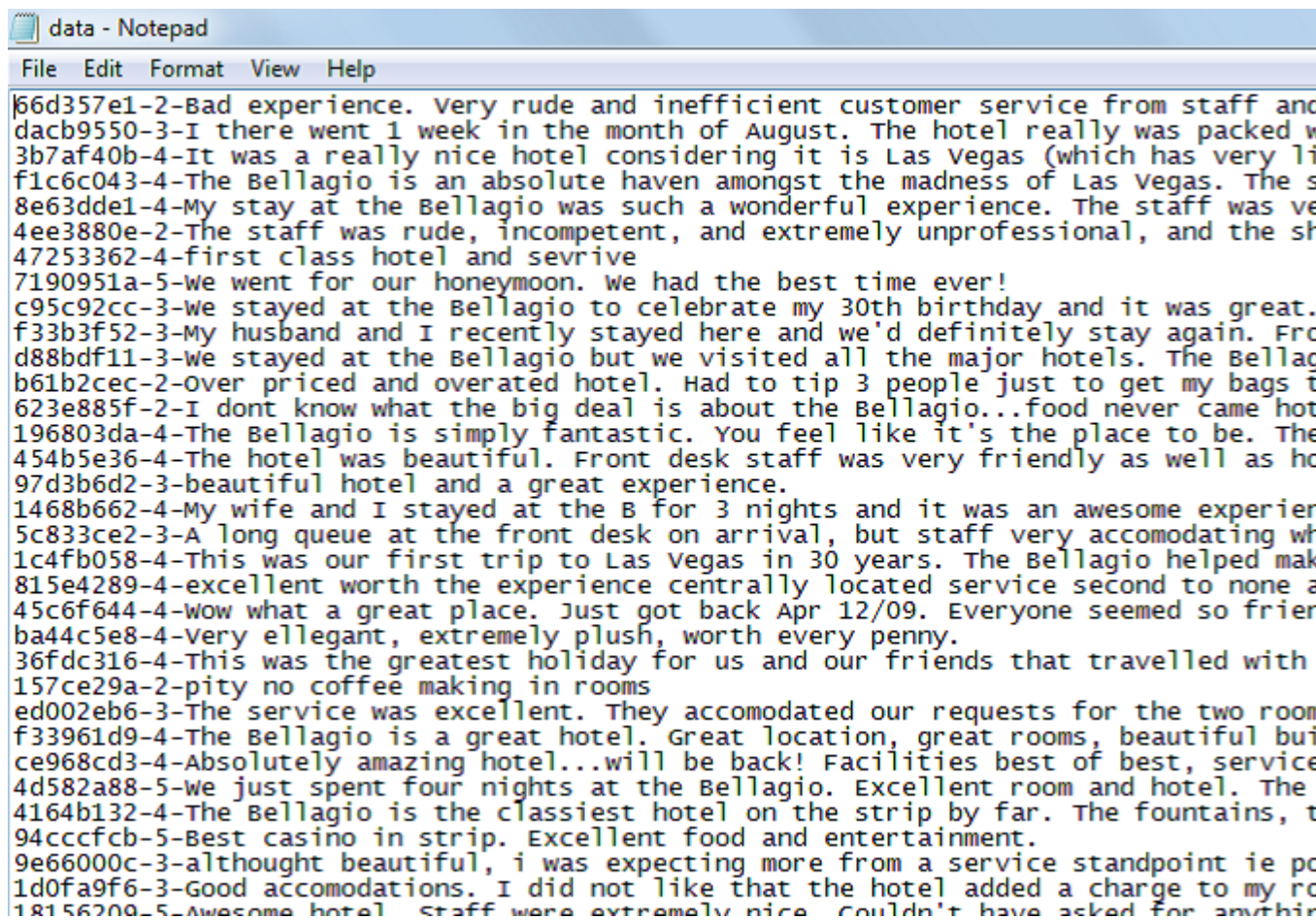


Figure 4.1: Collection of Hotel dataset

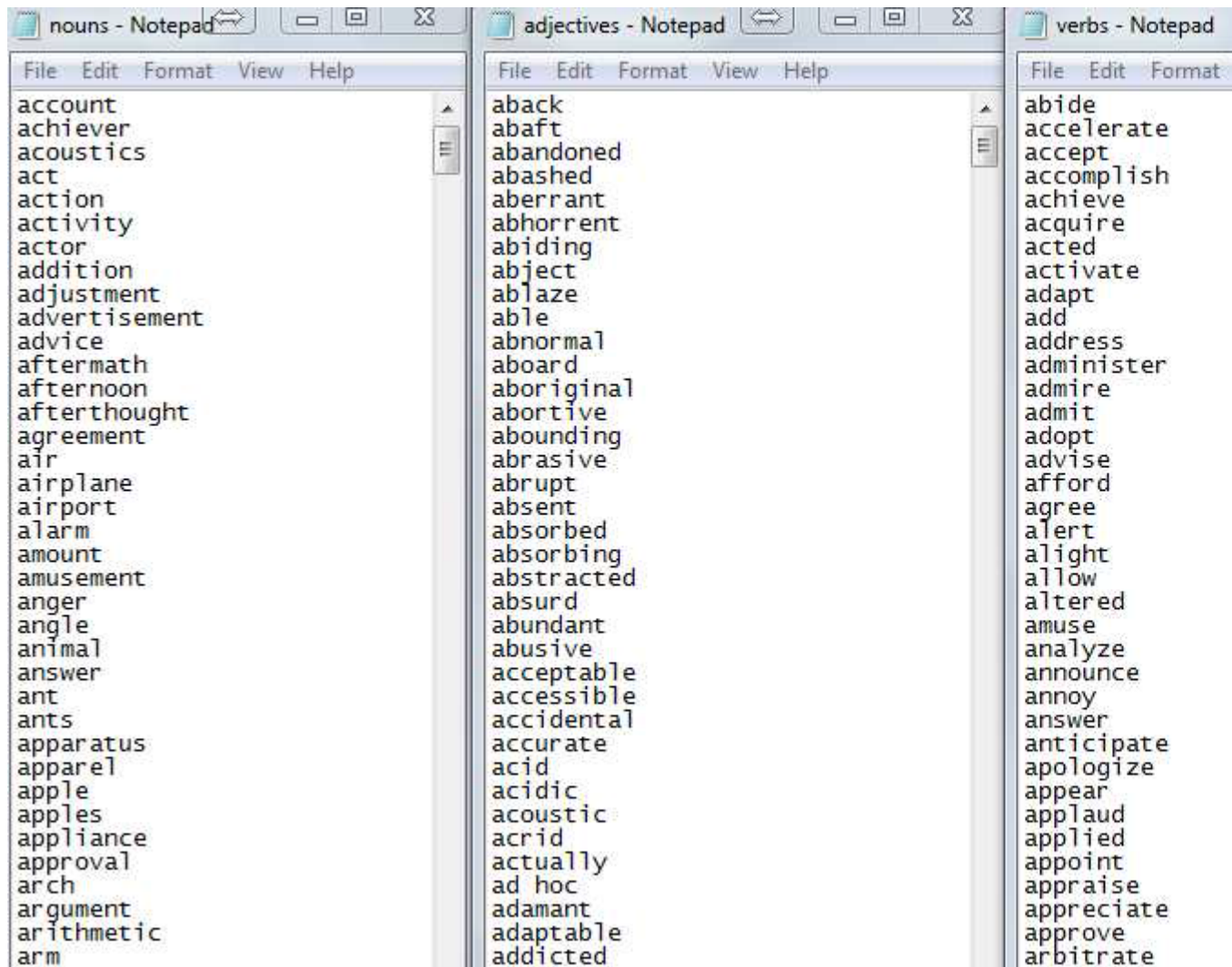


Figure 4.2: Extraction of n-gram

```

Python 2.7.5 Shell
File Edit Shell Debug Options Windows Help
Python 2.7.5 (default, May 15 2013, 22:43:36) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> ===== RESTART =====
>>>
hotel inefficient service staff bellagio nice wonderful beautiful experience air art
ny condition daughter day desk dinner door drink experience fact floor food front fu
onth morning night order part person place price regret room shop smell smoke star s
zing awesome best better big closed complex even excellent far first four free hard
l special strong superb two waiting whole young especially extremely not truly be be
love maintain make mistake order own pay perfect put queue recommend return say sche
>>> |

```

Figure 4.3: select data

## 4.1 dataset

text = """What can I say about this place. The staff of the restaurant is good and the eggplant is not bad. Apart from that, very uninspired food, lack of atmosphere and too expensive. I am a staunch vegetarian and was sorely dissapointed with the veggie options on the menu. Will be the last time I visit, I recommend others to avoid."""

```
Python 2.7.5 (default, Dec 10 2011, 12:21:33) [MSC v.1300 32
Type "copyright", "credits" or "license()" for more informat
>>> ===== RESTART =====
>>>
[['What', 'can', 'I', 'say', 'about', 'this', 'place', '.'],
 ['The',
  'staff',
  'of',
  'the',
  'restaurant',
  'is',
  'good',
  'and',
  'the',
  'eggplant',
  'is',
  'not',
  'bad',
  '.'],
 ['Apart',
  'from',
  'that',
  ',',
  'very',
  'uninspired',
  'food',
  ',',
  'lack',
  'of',
  'atmosphere',
  'and',
  'too',
  'expensive',
  '.'],
```

Figure 4.4: POS

And also calculate using genetics algorithm ,SVM ,and Entropy weights GA techniques to find the sentiment accuracy on movie reviews.

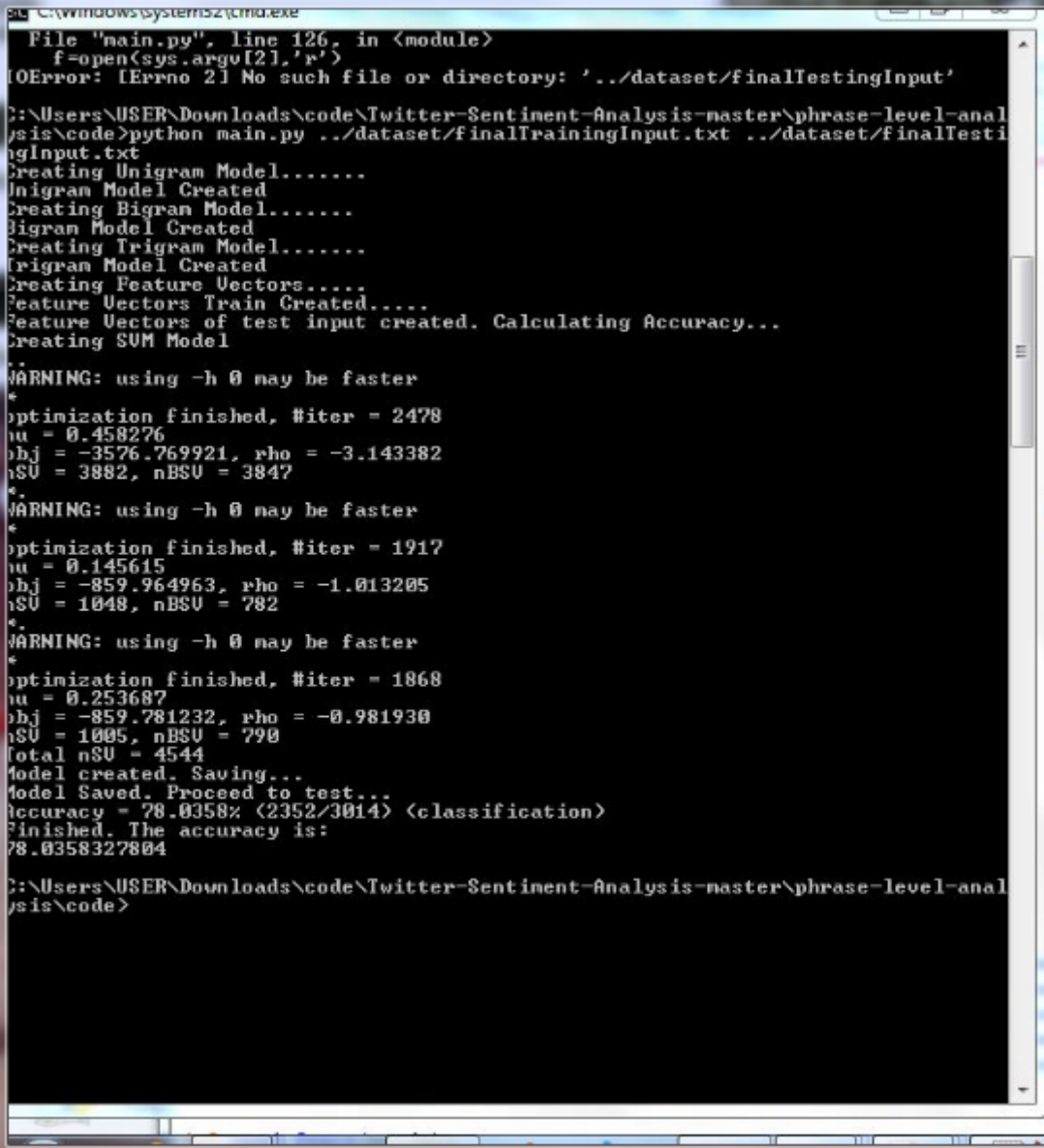
The dataset has been used for document level sentiment . The tested Consists

```
Python 2.7.9 (default, Dec 10 2014, 12:24:55) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> ===== RESTART =====
>>>
[-1.0, 11.0, 0.0, 1.0, 0.0, 1.0, 0.0, 3.0, -2.0, 1.0, 1.0, 5.0, 2.0, 1.0, 1.0, 4.0, -
positive = 14
negative = 3
neutral = 3
>>>
```

Figure 4.5: sentiment

of 2000 reviews (1000(+ve) and 1000 (-ve)) .The positive reviews are comprised of four and five star reviews archives while the negative ones are those receiving ones and two stars. An SVM was run using tenfold cross-validation ,with 1800 reviews used for training and 200 for testing in each fold and Bootstrapping was performed by randomly selecting 100 reviews for testing and remaining 1900 for training.

According to table figure 4.7 p-values for pair wise t-test on accuracy basis ,improved result using features selection and plot a graph.



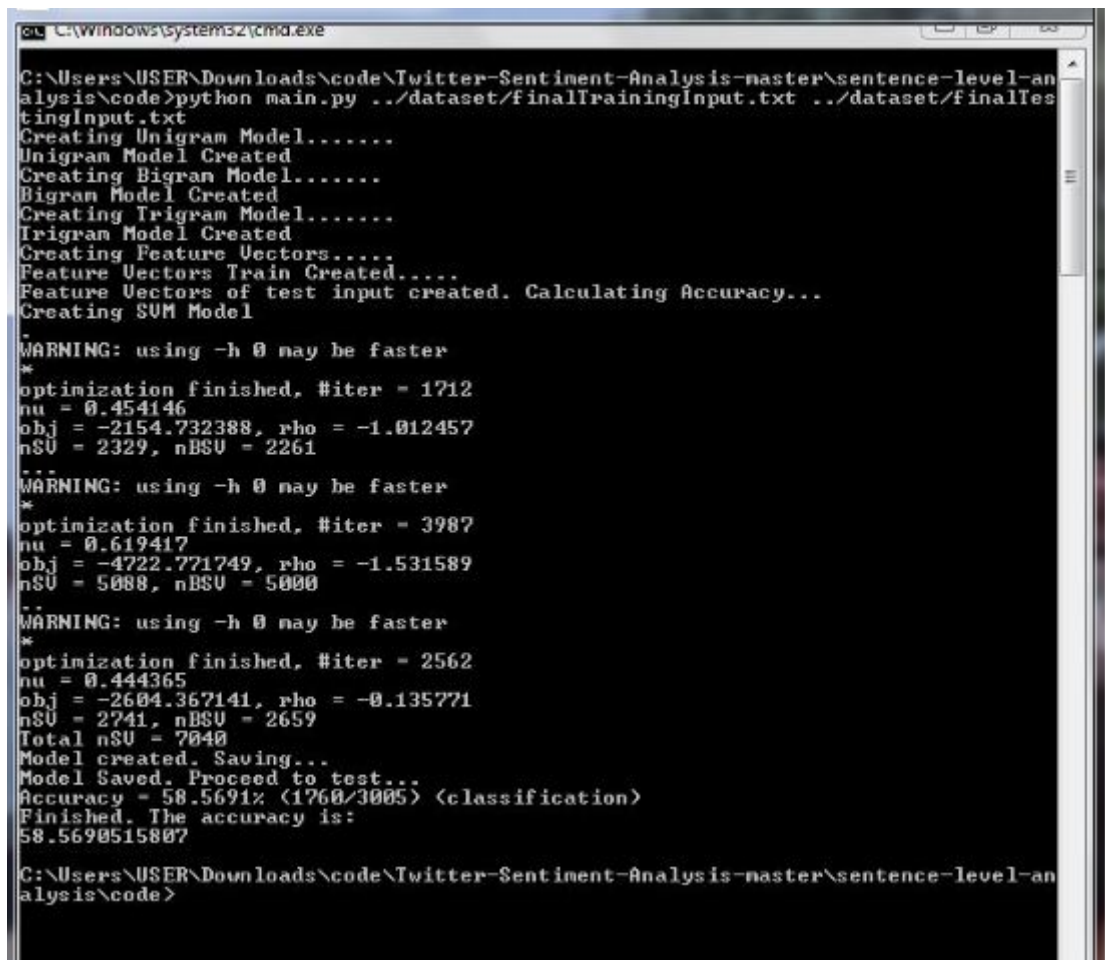
```
C:\windows\system32\cmd.exe
File "main.py", line 126, in <module>
  f=open(sys.argv[2], 'r')
IOError: [Errno 2] No such file or directory: '../dataset/finalTestingInput'

C:\Users\USER\Downloads\code\Twitter-Sentiment-Analysis-master\phrase-level-analysis\code>python main.py ../dataset/finalTrainingInput.txt ../dataset/finalTestingInput.txt
Creating Unigram Model.....
Unigram Model Created
Creating Bigram Model.....
Bigram Model Created
Creating Trigram Model.....
Trigram Model Created
Creating Feature Vectors.....
Feature Vectors Train Created.....
Feature Vectors of test input created. Calculating Accuracy...
Creating SUM Model
..
WARNING: using -h 0 may be faster
*
optimization finished, #iter = 2478
uu = 0.458226
ohj = -3576.769921, rho = -3.143382
nSU = 3882, nBSU = 3847
*
WARNING: using -h 0 may be faster
*
optimization finished, #iter = 1917
uu = 0.145615
ohj = -859.964963, rho = -1.013205
nSU = 1048, nBSU = 782
*
WARNING: using -h 0 may be faster
*
optimization finished, #iter = 1868
uu = 0.253687
ohj = -859.781232, rho = -0.981930
nSU = 1005, nBSU = 790
Total nSU = 4544
Model created. Saving...
Model Saved. Proceed to test...
Accuracy = 78.0358% (2352/3014) <classification>
Finished. The accuracy is:
78.0358327804

C:\Users\USER\Downloads\code\Twitter-Sentiment-Analysis-master\phrase-level-analysis\code>
```

Figure 4.6:





```
C:\Windows\system32\cmd.exe
C:\Users\USER\Downloads\code\Twitter-Sentiment-Analysis-master\sentence-level-an
alysis\code>python main.py ../dataset/finalTrainingInput.txt ../dataset/finalTes
tingInput.txt
Creating Unigram Model.....
Unigram Model Created
Creating Bigram Model.....
Bigram Model Created
Creating Trigram Model.....
Trigram Model Created
Creating Feature Vectors.....
Feature Vectors Train Created.....
Feature Vectors of test input created. Calculating Accuracy...
Creating SVM Model
-
WARNING: using -h 0 may be faster
*
optimization finished, #iter = 1712
nu = 0.454146
obj = -2154.732388, rho = -1.012457
nSV = 2329, nBSV = 2261
--
WARNING: using -h 0 may be faster
*
optimization finished, #iter = 3987
nu = 0.619417
obj = -4722.771749, rho = -1.531589
nSV = 5088, nBSV = 5000
--
WARNING: using -h 0 may be faster
*
optimization finished, #iter = 2562
nu = 0.444365
obj = -2604.367141, rho = -0.135771
nSV = 2741, nBSV = 2659
Total nSV = 7040
Model created. Saving...
Model Saved. Proceed to test...
Accuracy = 58.5691% (1760/3005) <classification>
Finished. The accuracy is:
58.5690515807

C:\Users\USER\Downloads\code\Twitter-Sentiment-Analysis-master\sentence-level-an
alysis\code>
```

Figure 4.7:

# Chapter 5

## Conclusion

We proposed a set of techniques for mining and summarizing product reviews based on data mining and natural language processing methods. To provide a features based summary of large number of customer reviews of a hotel, movie and experimental results indicates that the proposed techniques are very promising in performing their tasks. We believe that monitoring will be particularly useful to product in new positive or negative comments on their.



# Bibliography

- [1] Ahmed Abbasi, Hsinchun Chen, and Arab Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3):12, 2008.
- [2] Kerstin Denecke and Yihan Deng. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial Intelligence in Medicine*, 2015.
- [3] Pravesh Kumar Singh and Mohd Shahid Husain. Methodological study of opinion mining and sentiment analysis techniques. *International Journal on Soft Computing*, 5(1), 2014.
- [4] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. pages 168–177, 2004.
- [5] Anuj Sharma and Shubhamoy Dey. A comparative study of feature selection and machine learning techniques for sentiment analysis. pages 1–7, 2012.
- [6] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.