Improving Product rating system with text mining on product review/comments

April 30, 2014

Declaration

We do hereby that this Paper (Improving Product rating system with text mining on product review/comments) is submitted to Computer Science and Engineering Department of BRAC University as thesis, to complete Bachelor of Computer Science and Engineering Degree. Contents of the paper are original or otherwise provided with proper reference. This paper was not used for any other academic or non-academic purpose.

Md. Ahsanul Karim 10101023

Raduanul Islam 14141015

Sk. Wasif Ahmed 10101027

Supervisor: Prof. Md. Zahidur Rahman

Acknowledgment

We would humbly like to thank everyone who has helped in completion of this thesis work, for their advice, suggestion and help. We cordially thank our supervisor Professor Zahidur Rahman sir,our co-supervisor Abdur Rahman Adnan sir, for their support and helpful advise. We would like to thank our family and friends for their support and motivation.

Contents

List of Figures

Abstract

It is often considered a better exercise to have a complete idea of a particular service or a product before availing it. Now a days almost every online shopping sites or even the manufacturer of the product has a star based rating system and review/comment zone in their website. It is often not feasible to go through all the review before purchasing or availing that particular product. So people often tends to have an idea based on the number of stars on that product.Currently available systems use a star based rating where people rate the service or product on the scale of 5 or 10. The problem with that is when they give those rating stars they often tends to give it without giving much thought to it. User experience level and his mind set while rating varies very much. For example a person who loves particular brand of Soda, if he drink soda of another brand he might rate it lower than what it should be because he is used to a particular brand. But when he writes a review the chances are higher that he will write the major positive and negative aspects of that product . Thus the chances of getting a better feedback comes when it is review rather than stars. But as it has been mentioned earlier, going through all the reviews are not feasible. So we tried to improve the rating system by extracting information from the review text by using text mining technique upon that.

Chapter 1

Introduction

Text Mining is one of the upcoming field of interest for researchers around the world. We have used Text mining technique for analyzing product review to improve product rating system for our thesis. Currently available star based rating system does not always give proper information about any particular product. When people rate a particular product they often tends to do it casually without putting much thought about it. For example, a product may have 10 features, out of them 9 are working perfectly and 1 is not working properly, if some rate this 1 or 2 out of 10 it is not justified. On the other hand when a person writes a review or comment on the product he generally puts more thought on to it so extracting information out of that will be more helpful to rate the product properly.

1.1 Data Set

1.1.1 Data Collection

We have used a data set consisting over 500000 reviews of product sold on amazon.com.The data set contains product rating and reviews . this data set was used both for training and testing purpose.

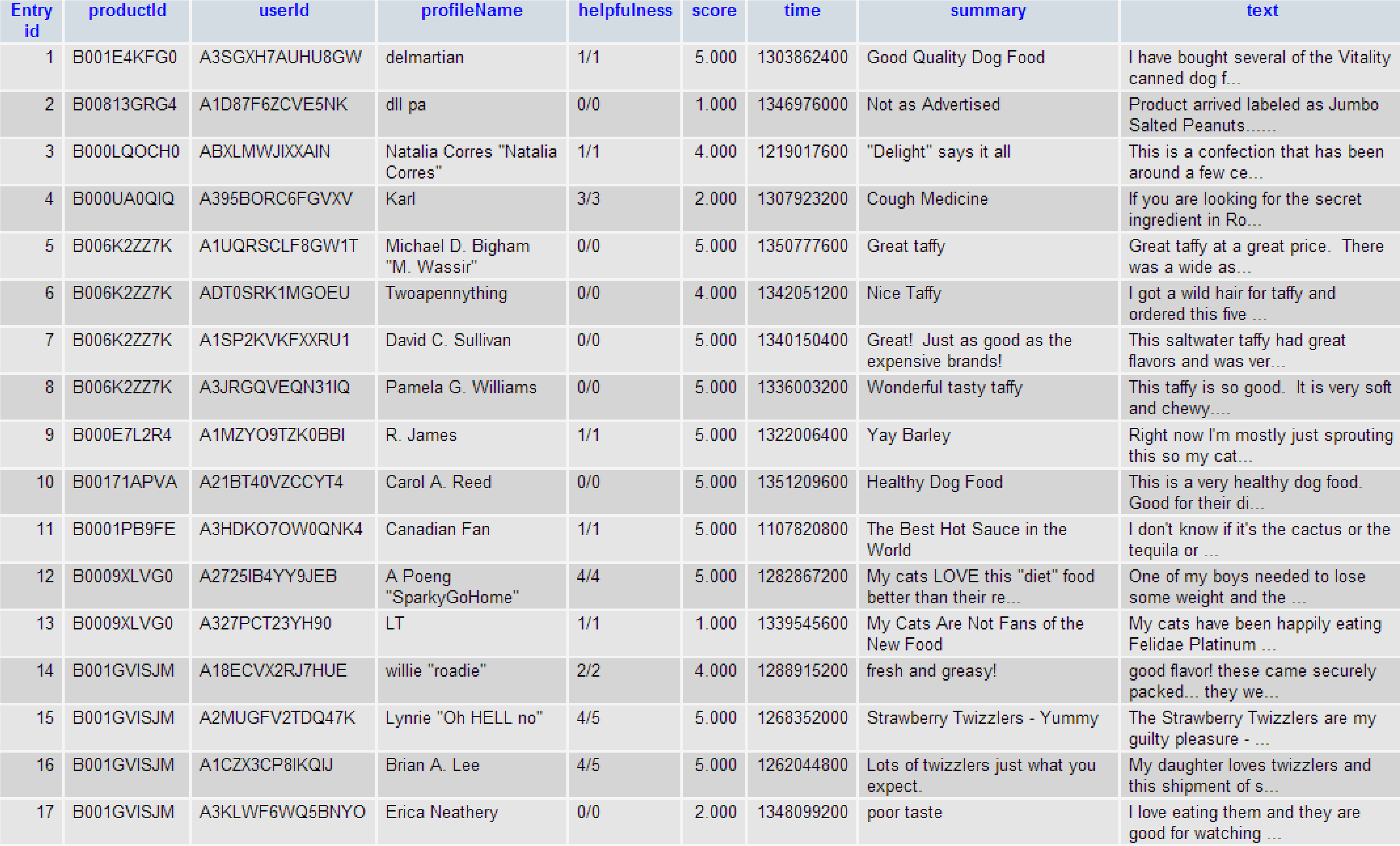
Data Statistic

Number of reviews	568,454
Number of users	256,059
Number of products	74,258
Users with > 50 reviews	260
Median no. of words per review	56
Time span	Oct 1999 - Oct 2012

Data format product/product Id: B001E4KFG0
review/userId: A3SGXH7AUHU8GW
review/profileName: delmartian
review/helpfulness: 1/1
review/score: 5.0
review/time: 1303862400
review/summary: Good Quality Dog Food review/text: I have bought several of the Vitality canned dog food products and have found them all to be of good quality. The product looks more like a stew than a processed meat and it smells better. My Labrador is finicky and she appreciates this product better than most.

1.2 Brief summary of the chapters to come

At First data prepossessing is done. Program written on C++ and Python was used for data extraction , punctuation and stop words handling stemming , hashing and frequency counting. Once the data was Preprocessed it was run on Octave using Support vector machine .



Entry id	productId	userId	profileName	helpfulness	score	time	summary	text
1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1/1	5.000	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned dog f...
2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0/0	1.000	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanuts......
3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1/1	4.000	1219017600	"Delight" says it all	This is a confection that has been around a few ce...
4	B000UA0QIQ	A395BORC6FGVXV	Karl	3/3	2.000	1307923200	Cough Medicine	If you are looking for the secret ingredient in Ro...
5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0/0	5.000	1350777600	Great taffy	Great taffy at a great price. There was a wide as...
6	B006K2ZZ7K	ADT0SRK1MGOEU	Twoapennything	0/0	4.000	1342051200	Nice Taffy	I got a wild hair for taffy and ordered this five ...
7	B006K2ZZ7K	A1SP2KVKFXXRU1	David C. Sullivan	0/0	5.000	1340150400	Great! Just as good as the expensive brands!	This saltwater taffy had great flavors and was ver...
8	B006K2ZZ7K	A3JRGQVEQN31IQ	Pamela G. Williams	0/0	5.000	1336003200	Wonderful tasty taffy	This taffy is so good. It is very soft and chewy....
9	B000E7L2R4	A1MZYO9TZK0BBI	R. James	1/1	5.000	1322006400	Yay Barley	Right now I'm mostly just sprouting this so my cat...
10	B00171APVA	A21BT40VZCCYT4	Carol A. Reed	0/0	5.000	1351209600	Healthy Dog Food	This is a very healthy dog food. Good for their di...
11	B0001PB9FE	A3HDKO7OW0QNK4	Canadian Fan	1/1	5.000	1107820800	The Best Hot Sauce in the World	I don't know if it's the cactus or the tequila or ...
12	B0009XLVG0	A2725IB4YY9JEB	A Poeng "SparkyGoHome"	4/4	5.000	1282867200	My cats LOVE this "diet" food better than their re...	One of my boys needed to lose some weight and the ...
13	B0009XLVG0	A327PCT23YH90	LT	1/1	1.000	1339545600	My Cats Are Not Fans of the New Food	My cats have been happily eating Felidae Platinum ...
14	B001GVISJM	A18ECVX2RJ7HUE	willie "roadie"	2/2	4.000	1288915200	fresh and greasy!	good flavor! these came securely packed... they we...
15	B001GVISJM	A2MUGFV2TDQ47K	Lynrie "Oh HELL no"	4/5	5.000	1268352000	Strawberry Twizzlers - Yummy	The Strawberry Twizzlers are my guilty pleasure - ...
16	B001GVISJM	A1CZX3CP8IKQIJ	Brian A. Lee	4/5	5.000	1262044800	Lots of twizzlers just what you expect.	My daughter loves twizzlers and this shipment of s...
17	B001GVISJM	A3KLWF6WQ5BNYO	Erica Neathery	0/0	2.000	1348099200	poor taste	I love eating them and they are good for watching ...

Figure 1.1: Data Set View Using Xampp

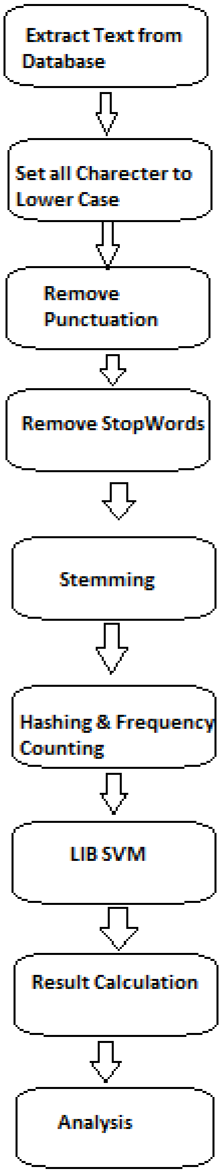

Extract Text from Database
Set all Charecter to Lower Case
Remove Punctuation
Remove StopWords
Stemming
Hashing & Frequency Counting
LIB SVM
Result Calculation
Analysis

Figure 1.2: Data Processing Flow Chart

Chapter 2

Related Work

In this paper[5] Durgesh and Lekha used some data set and applied different types of kernels and showed the accuracy level. Here they selected rough data sets (e.g. high dimensions) and also they predicted the class lavel and showed the time svm and RSES(tool set for data classification, using different classifier technique as Rule Based classifier, Rule Based classifier with Discretization, K-NN classifier and LTF (Local Transfer Function) Classifier) has taken to predict the level and at last they compare the accuracy svm and RSES.

These paper[1] did text categorization using svm and Joachim(2009) have used polynomial, and RBF Kernels and did binary classification also there were different data sets. For every data set the showed the accuracy lavel. It is also found that the result of other commonly used learning methods for text categorization and their results show that for text categorization using svm, accuracy level is much higher than the accuracy level using learning methods for text categorization[9]. On the other hand we find out the accuracy over one data set and also predicted the class label for any review.

Chapter 3

Data Prepossessing

A series of steps are followed while processing the data set . All the steps of data processing are discussed briefly below

3.1 Extracting Information from Data set

We have used Product review Data set of Amazon.com. This data set contains product id ,user id , profile name ,helpfulness , score, time ,summery and text(comment/review). 'Product id' represents a unique number which represents a particular Product , 'user id' is randomly yet uniquely generated by the website, 'profile name' is set by the person who is giving the review,'helpfulness' represents how many people has found it to be helpful, 'score' is the rating out of 5 scale,' time' is the time of review in UNIX time , 'summary' represents the summary of the review by the user/ reviewer and than comes the actual review(text) . A simple program on C++ language was written to extract only the text(comment/review) , this is the only part of the data set which we have used further for our thesis.

3.2 Changing Character case

We set all the character in the review text to lower case. It helped us to avoid redundant text . for example : Good, GOOD, good all of these 3 means the same but due to different character case they appeared as separate entity and thus results unwanted calculation, wasting both time and space .

3.3 Removing Punctuation

All punctuation (e.g. , . ? !) are discarded afterward. In this system punctuations were completely avoided during frequency generation.

3.4 Removing Stop word

In this paper Prepositions (e.g.at,of,for,to,after), conjunctions (e.g.:and,or,but) and article (a,an, the) are considered as stop words. These stop words could cause unwanted increase of frequency resulting computational complication thus reduce performance.

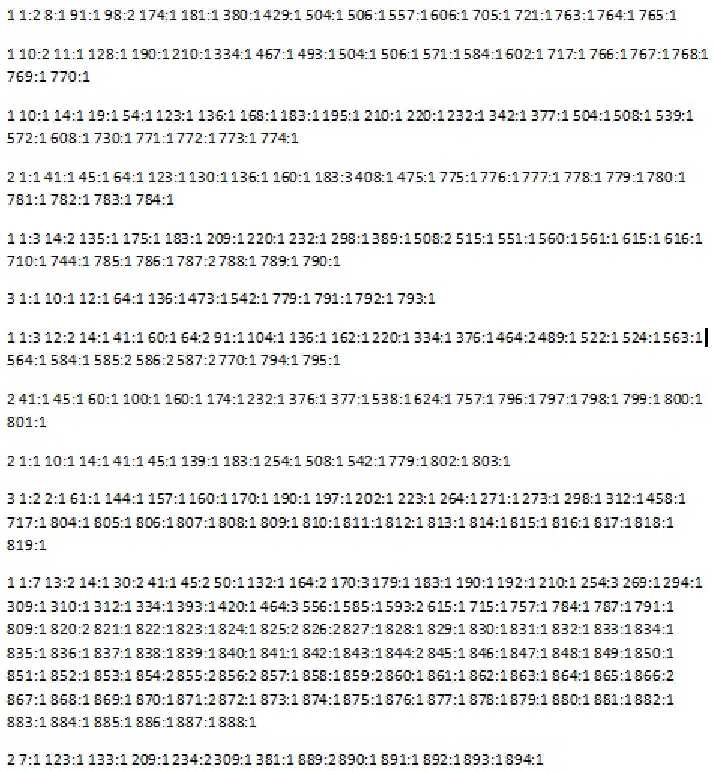
3.5 Stemming

In linguistic morphology and information retrieval, stemming is the process for reducing inflected (or derived) words to their stem, base or root form generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. For example the root word 'eat' can be used as 'eating', 'eaten', or even 'ate'.

A program was written in Python Language was used for stemming,changing character case, removing punctuation and stop words.

3.6 Hashing and frequency Counting

After stemming all the key words from the review were hashed which gives unique id for each word. Then for each review, occurrence or those key words were added up and frequency of each word was generated.



```
1 1:2 8:1 91:1 98:2 174:1 181:1 380:1429:1 504:1 506:1557:1606:1 705:1 721:1763:1764:1 765:1

1 10:2 11:1 128:1 190:1210:1334:1 467:1 493:1504:1 506:1 571:1584:1602:1 717:1 766:1767:1768:1
769:1 770:1

1 10:1 14:1 19:1 54:1 123:1 136:1 168:1 183:1 195:1 210:1 220:1 232:1 342:1 377:1 504:1508:1 539:1
572:1 608:1 730:1 771:1772:1773:1 774:1

2 1:1 41:1 45:1 64:1 123:1 130:1 136:1 160:1 183:3408:1 475:1 775:1776:1777:1 778:1 779:1780:1
781:1 782:1 783:1 784:1

1 1:3 14:2 135:1 175:1 183:1 209:1 220:1 232:1 298:1 389:1508:2 515:1 551:1560:1561:1 615:1 616:1
710:1 744:1 785:1 786:1787:2 788:1 789:1 790:1

3 1:1 10:1 12:1 64:1 136:1473:1542:1 779:1 791:1792:1 793:1

1 1:3 12:2 14:1 41:1 60:1 64:2 91:1 104:1 136:1 162:1 220:1 334:1 376:1464:2489:1 522:1 524:1563:1
564:1 584:1 585:2 586:2587:2770:1 794:1 795:1

2 41:1 45:1 60:1 100:1 160:1 174:1232:1 376:1 377:1538:1624:1 757:1 796:1797:1798:1 799:1 800:1
801:1

2 1:1 10:1 14:1 41:1 45:1 139:1 183:1254:1 508:1 542:1779:1802:1 803:1

3 1:2 2:1 61:1 144:1 157:1 160:1 170:1 190:1 197:1 202:1 223:1 264:1 271:1 273:1 298:1 312:1 458:1
717:1 804:1 805:1 806:1807:1808:1 809:1 810:1811:1812:1 813:1 814:1815:1 816:1 817:1818:1
819:1

1 1:7 13:2 14:1 30:2 41:1 45:2 50:1 132:1 164:2 170:3 179:1 183:1 190:1 192:1 210:1 254:3 269:1 294:1
309:1 310:1 312:1 334:1393:1420:1 464:3 556:1585:1593:2 615:1 715:1757:1 784:1 787:1 791:1
809:1 820:2 821:1 822:1823:1824:1 825:2 826:2827:1828:1 829:1 830:1831:1 832:1 833:1834:1
835:1 836:1 837:1 838:1839:1840:1 841:1 842:1843:1844:2 845:1 846:1847:1 848:1 849:1850:1
851:1 852:1 853:1 854:2855:2856:2 857:1 858:1859:2860:1 861:1 862:1863:1 864:1 865:1866:2
867:1 868:1 869:1 870:1871:2872:1 873:1 874:1875:1876:1 877:1 878:1879:1 880:1 881:1882:1
883:1 884:1 885:1 886:1887:1888:1

2 7:1 123:1 133:1 209:1234:2309:1 381:1 889:2890:1 891:1 892:1893:1894:1
```

Figure 3.1: View after hashing

Chapter 4

Structure of solution

4.1 Support vector machine (SVM)

SVM has gain significant popularity in the field of text classification over state-of-the art-methods. SVM is very robust and it eliminates the need of expensive parameter tuning[1].Support vector machines are based on the Structural Risk Minimization principle from computational learning theory. The idea of structural risk minimization is to find a hypothesis h for which we can guarantee the lowest true error. The true error of h is the probability that h will make an error on an unseen and randomly selected test example. An upper bound can be used to connect the true error of a hypothesis h with the error of h on the training set and the complexity of H (measured by VC-Dimension), the hypothesis space containing h. Support vector machines and the hypothesis h which (approximately) minimizes this bound on the true error by selectively and efficiently controlling the VC-Dimension of H.[1]

4.1.1 SVMs Work Well for Text Categorization

To Find out what methods are promising for learning text classifiers, we should Find out more about the properties of text

.

High dimensional input space: When learning text classifiers, one has to deal with many (more than 10000) features. Since SVMs use overtting protection, which does not necessarily depend on the number of features, they have the potential to handle these large feature spaces.

Few irrelevant features: One way to avoid these high dimensional input spaces is to assume that most of the features are irrelevant. Feature selection tries to determine these irrelevant features. Unfortunately, in text categorization there are only very few irrelevant features.

Document vectors are sparse: For each document, the corresponding document vector contains only few entries which are not zero. Kivinen et al. [3] give both theoretical and empirical evidence for the mistake bound model that

\additive" algorithms, which have a similar inductive bias like SVMs, are well suited for problems with dense concepts and sparse instances.

Most text categorization problems are linearly separable: All Ohsumed categories are linearly separable and so are many of the Reuters tasks. The idea of SVMs is to find such linear (or polynomial, RBF, etc.) separators. These arguments give theoretical evidence that SVMs should perform well for text categorization.[1]

SVMs consistently achieve good performance on text categorization tasks, outperforming existing methods substantially and significantly.With their ability to generalize well in high dimensional feature spaces, SVMs eliminate the need for feature selection, making the application of text categorization considerably easier. Another advantage of SVMs over the conventional methods is their robustness. SVMs show good performance in all experiments, avoiding catastrophic failure, as observed with the conventional methods on some tasks. Furthermore, SVMs do not require any parameter tuning, since they can find good parameter settings automatically. All this makes SVMs a very promising and easy-to-use method for learning text classifier from examples.

4.1.2 SVM Kernel

Out of several SVM kernel we worked on RBF Kernel and Linear Kernel.

4.1.2.1 SVM LIB

The main Reason for choosing SVM LIB is that it is very easy to work with and it is very good for classifications purposes. LIBSVM is an integrated software for support vector classification, (C-SVC, nu-SVC), regression (epsilon-SVR, nu-SVR) and distribution estimation (one-class SVM). It supports multi-class classification. It helps to easily use SVM as a tool. LIBSVM provides a simple interface where users can easily link it with their own programs. Main features of LIBSVM include

- Different SVM formulations
- Efficient multi-class classification
- Cross validation for model selection
- Probability estimates
- Various kernels (including precomputed kernel matrix)
- Weighted SVM for unbalanced data
- Both C++ and Java sources
- GUI demonstrating SVM classification and regression
- Python, R, MATLAB, Perl, Ruby, Weka, Common LISP, CLISP, Haskell, OCaml, LabVIEW, and PHP interfaces. C# .NET code and CUDA extension is available. It's also included in some data mining environments: RapidMiner, PCP, and LIONsolver.

Data Format:: The format of training and testing data file is:

<label><index1>:<value1><index2>:<value2> ......

Label is a integer indicating class,value which can be any real number. For one-class SVM, it's not used so can be any number. The pair <index>:<value> gives a feature (attribute) value: <index> is an integer starting from 1 and <value>. So we have created our data as svm data format. Now the 2nd procedure will be selecting a kernel for classification. There are four common kernels , we must decide which one to try first. Then the penalty parameter C and kernel parameters are chosen.

4.1.2.2 SVM Kernel results

In general, the RBF kernel is a reasonable first choice. This kernel is non linearly maps samples into a higher dimensional space so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is non-linear. Furthermore, the linear kernel is a special case of RBF Keerthiand Lin (2003)[7] since the linear kernel with a penalty parameter $C^{\sim}$ has the same performance as the RBF kernel with some parameters (C; gaama). In addition, the sigmoid kernel behaves like RBF for certain parameters.(Lin and Lin, 2003)[7].We tried to run on both kernels and came up with following results.

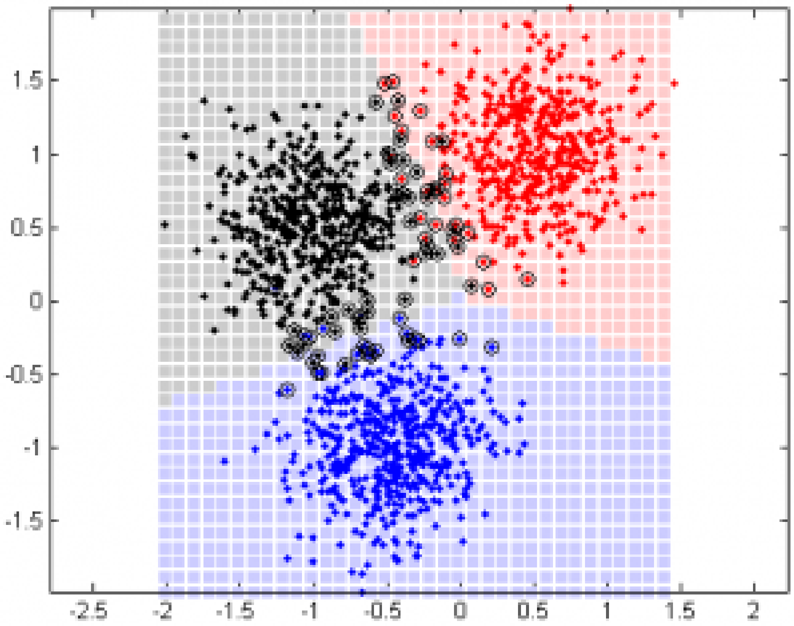
For our training we used 3000 training data and 1000 test data. As our Data Set was huge and their was hardware constrains like limited main memory and processing speed we ran chunk of 50,000 data each time on the mechine.

1-50,000————->78%
50,001-100,000—–>77.65%
150,001-200,000—–>79.2%
200,001-250,000—–>71%
250,001-350,000—–>70.3%
350,001-568,450—–>66.3%
Results omitting after 2 decimal point.

Result of RBF Kernel:
Accuracy(Class 1) =76.0067% (For Positive review)
Accuracy(Class 2)=74.6% (For negative Review)
Accuracy(Class 3)=96.933% (For neutral Review)
Accuracy=73.8667% (Total Accuracy)

Linear Kernel::: Accuracy(Mean)=82.22%

Figure 4.1: Linear kernel using 3 classes



Chapter 5

Data Classification and Analysis

5.1 Text Review Analysis

To understand trends in reviewing behaviors, we performed an in-depth analysis of 568450 user review augmented with our automatic classification. Thus, we could study the relation between the textual structure of the reviews and the meta data entered by the reviewers, such as star rating.

5.2 User review Trends

Our analysis of the reviews shows that the sentiment expressed in the reviews was mostly positive.(About 75% of the reviews were positive, 18% were negative and only 7% were neutral) This is consistent with the star rating provided by users, with 78% of reviews having a star rating of 4 or 5.

Most reviews describe the food/product quality and the amazon service to provide the goods to the reviewer. In the negative review most reviews complains about the shipping service and food quality. But mostly about shipping service. And in neutral review mostly about amazon shipping service.

5.3 Average Metadata Based Prediction

In our classification we saw that there were a number of mismatch between user rating and user review. We strongly suggest that product rating should be calculated using product review. We have used 4 or 5 as positive 3 neutral and 1 or 2 rating as negative rating . We believe that with this rating system , the product rating will be more realistic.

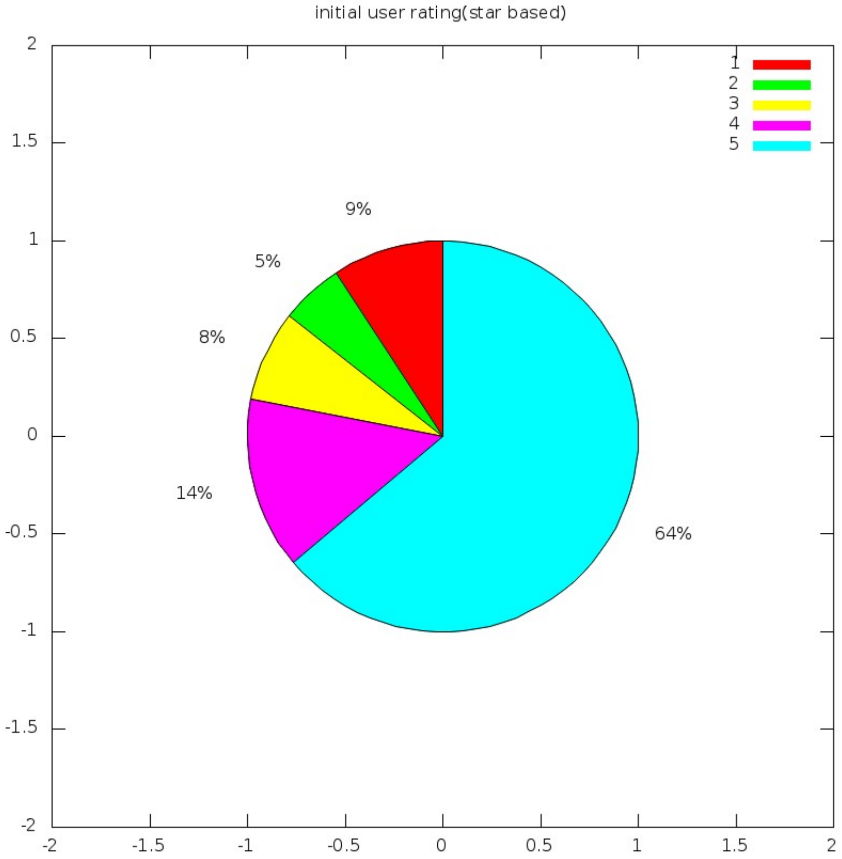


Figure 5.1: Star Based System rating

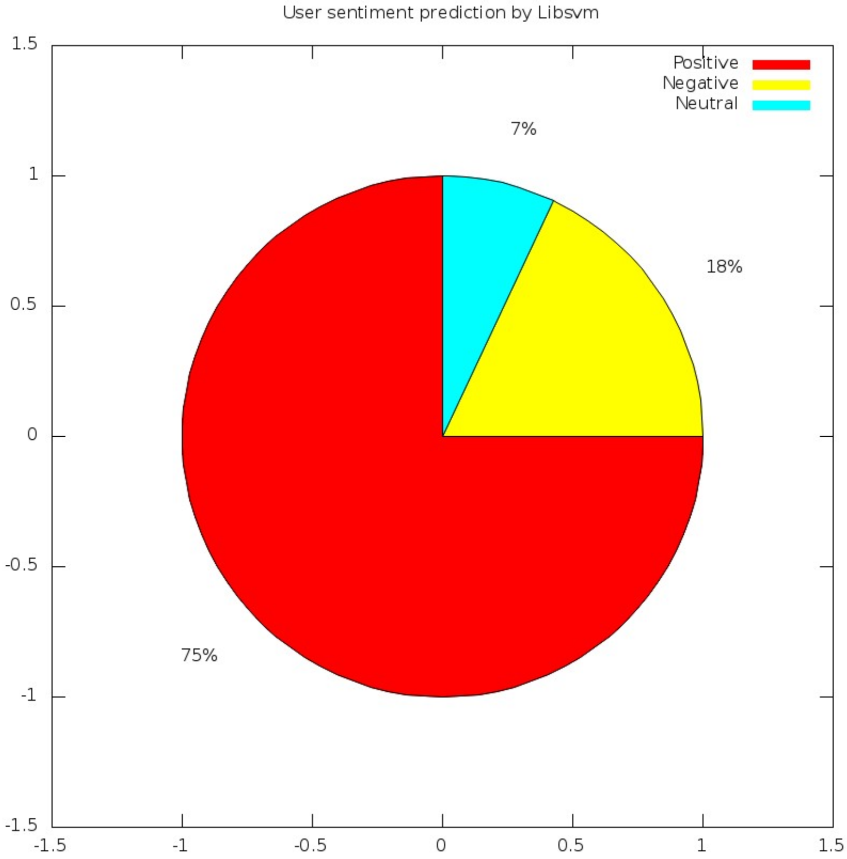


Figure 5.2: Sentiment Analysis rating

Chapter 6

Conclusion and Future Works

In this paper we have studied about Text mining technique on Amazon.com's Food Review database.It was quite a big data set and there were a huge range of review for different products. The accuracy we got, varies because of the range of review and product.In Future we would like to do the following things

- user's experience level analysis to get more accurate review
- Recommend related product to the Customers/users
- Optimize searching technique to ensure most related and helpful product feedback and review

Bibliography

[1] Joachims Thorsten(1998)Text Categorization with Support Vector Machines:Learning With Many Relevant Features.

[2] Julian McAuley,Jure Leskovec(2013)From Amateur to Connoisseur Modeling the Evolution of User Expertise.

[3] J. Kivinen, M. Warmuth, and P. Auer(1995). The perceptron algorithm vs. winnow: Linear vs. logarithmic mistake bounds when few input variables are relevant. In Conference on Computational Learning Theory.

[4] Simon Tong, Daphne Koller (2001)Support Vector Machine Active Learning with Applications to Text Classification.

[5] Durgesh K. Srivastava, Lekha Bhambu(2009)Data Classification Using Support Vector Machine

[6] Yohan Jo,Alice Oh(2011)Aspect and Sentiment Unification Model for online Review Analysis

[7] Chih-Wei Hsu,Chih-Chung Chang,Chih-Jen Lin(2003) A Practical Guide to Support Vector Classification

[8] Miquing Hu, Bing Liu [KDD-04] Mining and Summarizing Customer Reviews.

[9] Gayatree Ganu, Noemie Elhadad, Amelie Marian(2009) Beyond the Stars: Improving Rating Predictions using Review Text Content.