Detection and Summarization of Genuine Review using Visual Data Mining

Jagruti Prajapati
Department of Information Technology
CHARUSAT, Changa, India

Malay Bhatt
Department of Computer Engineering
DDU, Nadiad, India

Dinesh J. Prajapati
Department of Information Technology
ADIT,V. V. Nagar, India

ABSTRACT
In earlier days we were asking our friends or relatives for their opinions regarding products which we want to purchase from the merchants. But now a day's E-commerce is gaining more and more popularity. Whatever query we are having, we can find its answer from World Wide Web. Merchants are also selling their products online and at a same time they are asking customer's review regarding products, which customer has bought. This would be beneficial to merchants as well as customers also. As the numbers of customers are growing, reviews received by products are also growing in large amount. Thus, mining opinions from product reviews is an important research topic. However, existing research is more focused towards classification and summarization of such online opinions. An important issue related to the trustworthiness of online opinions has been neglected most often. There is no reported study on assessing the trustworthiness of reviews. This research paper aims to first classify the opinion (positive or negative) carried out by detection of a review(spam or a non-spam) based on rating behavior and finally removing spam reviews, which provides a trusted review to help the customer in taking appropriate buying decision. This paper proposes a novel and effective technique, which will represent classified opinion in form of "chernoff face".

General Terms
Opinion mining, Visual data mining

Keywords
Chernoff face, Natural language processing, Opinion mining, Opinion spam, POS tagging, Semantic orientation, Spam detection

1. INTRODUCTION
As today World Wide Web is expanding very rapidly, E-commerce is also becoming very popular. Hence, more and more products sold online. As a result number of online customer also increases. In order to enhance the product quality and customer satisfaction, it has become a common practice for merchants, to ask their customers to give opinion on the products that they have purchased [5]. With the use of this practice, large number of customers are becoming familiar with products and they can conclude from the past customer's reviews/ opinions, whether to purchase a particular product or not? Thus, online reviews are helpful to both customer and merchants. As numbers of online customers are increasing, a review that a product receives also increases very rapidly. Furthermore, for an end customer it would be difficult to read all reviews and make decision of whether to purchase the product or not? Thus it is essential to have summarization of all reviews, which helps a customer to take appropriate decision. But this practice of asking customer for their reviews, gives good chances for "review spam" as anyone can write anything on web. Review spam refers writing fake reviews in order to promote or de-promote particular product. This deliberately misleads the potential customers. Detecting a spam reviews is a very critical task for opinion mining. This paper aims to detect a spam review, removal of spam reviews and representation of genuine reviews.

Figure 1: Genuine review [12]

Our task is divided in three parts. *1. Classification of review (positive or negative):* Review contains more than one opinion sentence. For extracting opinion words from opinion sentence we have used Stanford NLP parser. Semantic orientation of opinion words is calculated using algorithm given in [9]. Based on semantic orientation (SO) value classification of review is done, If SO value is positive then opinion is classified as positive and if value is negative opinion is classified as negative. *2. Classification of review (spam or non-spam):* Review has two main parts *1.Content* and 2. *Rating.* Rating is very important in opinion. Different opinion spam detection techniques are mentioned in below section. For understanding rating importance let's take one example. In fig.1 reviews for sony cyber-shot DSD-WS30 is taken from [12].

Total 183 reviews are there from amazon, average rating is 4 stars. Fig. 1 gives most useful review with high rating. And in fig.2 possible spam review is given, we can call it as spam because there is very much deviation in rating. There are 14 more reviews which gives 2 star rating but their content gives genuine opinion. Thus from given example we can conclude that rating is very important for classifying opinions as spam or non-spam. After classification, removal of spam review should be done. *3. Summarization of genuine reviews:* An icon based visual data mining technique "chernoff face" is used for visualizing genuine opinions, Dimensions are mapped to the

properties of a face. Chernoff face is shown in fig. 3 and properties of chernoff face are listed in table 1.

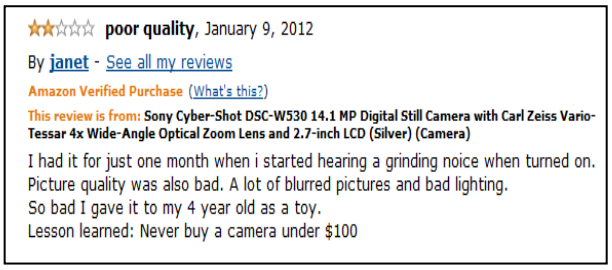


Figure 2: Spam review [12]

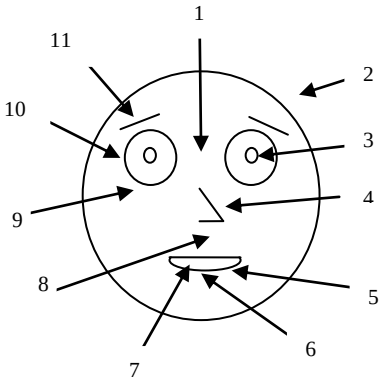


Figure 3: Chernoff face [18]

2. RELATED WORK

In this section, we survey the related research in "opinion mining and summarization" and "spam detection". In past enough work has been done for classifying opinion as positive or negative and summarizing [9]. For classifying opinion there are mainly three techniques [11]. *1. Sentiment classification* (document level classification): In document level, Turney [10] proposed an approach of determining document level classification by calculating the average semantic orientation (SO) of extracted phrases. SO was computed using PMI-IR algorithm. By using point wise mutual information (PMI) to measure the dependence between extracted phrases and the reference words "excellent" and "poor" and query is fired in search engine and hit counts are used for SO. *2. Feature based opinion mining and summarization* (sentence level classification): In sentence level Hu and Liu [9] proposed set of techniques for mining and summarizing product reviews. With the use of adjective synonym and antonym set in WordNet [16] Semantic Orientation (SO) is found for each opinion sentence. Then featured based summary is generated using high frequency feature words (the top ranked features) and ignoring infrequent features. *3. Comparative sentence and relation mining*. In this paper for classification of opinion (positive or negative) we have selected feature based opinion mining and summarization, as we want to summarize opinion using "chernoff face". Second we focus on spam detection. For detecting opinion Spam there are mainly three techniques [11], *1.Review centric spam detection:* In this approach, spam detection is based only on reviews. Review has two main parts: content and rating. We can detect possible spam activities based on content similarity and rating deviation. In [7] Jindal and Liu discovered spam Activities which are widespread. For example, they found a large number of duplicate and near-duplicate reviews written by the same reviewers on different products or by different reviewers (possibly different userids of the same persons) on the same products or different products. It makes the first attempt to investigate opinion spam in reviews and proposes some novel techniques to study spam detection [11]. In this paper for spam detection we have selected review centric spam detection with rating deviation.

Table 1: Properties of chernoff face

Feature no.	Name	Feature no.	Name
1	Eye spacing	7	Mouth openness
2	Head eccentricity	8	Nose width
3	Pupil size	9	Eye size
4	Nose length	10	Eye eccentricity
5	Mouth curvature	11	Eyebrow slope
6	Mouth width		

2. Reviewer centric spam detection: In this approach, "unusual" behaviors of reviewer are exploited for spam detection. In [3] Lim and Nguyen proposed a scoring method for detecting review spammers using rating behaviors. Then select a sub- set of highly suspicious reviewers for further scrutiny with the help of web based spammer evaluation software specially developed for user evaluation experiments. In [4] Jindal and Liu have considered spam detection as classification problem. For finding unexpected rules first, expected rules found then confidence unexpectedness and support unexpectedness is found for both one-conditional rules and three-conditional rules. These rules represent unusual behavior of reviewers, which indicates spam activities. *3. Server centric spam detection:* In this approach, server log at the review site can be helpful in spam detection. Another related research area is the group spam detection, a spammer *group* refers to a group of reviewers who Works together writing fake reviews to promote or demote a set of target products. Spammer groups are very damaging due to their sheer sizes [11]. When a group is working collaboratively towards a product, it can take control of the sentiment for the product. In [1] Mukherjee and Liu proposed a method to detect such groups, which consists of pattern mining to find candidate groups, assessing those using 8 criteria that indicate a typical behaviors of groups, and finally ranking the candidate groups using SVM ranking . Their experiment is based on a large set of Amazon reviewers and their reviews.

3. PROPOSED SYSTEM

Fig.4 illustrates the architecture of the proposed opinion spam detection system. Crawled (downloaded) dataset of product is supplied as input to the system. Summarization of genuine reviews is generated as output in form of "chernoff face". The system performs three main steps (as discussed in introduction):*1.Classification of review (positive or negative) 2.Detection of review (spam or non-spam).*

3. Summarization of genuine reviews. These steps are performed in multiple sub-steps. Description of each sub-step is given below.

3.1 Crawl reviews

The system first downloads (or crawls) all the reviews, and put them in the review database. For this research work dataset used is reviews from amazon.com [12, 13]. We are using this dataset because, Amazon.com is very popular and successful e-

commerce website for gathering reviews of customers. Also it contains listing of very wide range of products. Thus, it will be easy for customer to search for reviews for particular product.

3.2 Pre-processing of dataset

The dataset contains many opinions and on pre-processing an individual text file is generated for each opinion. Thus, there is one input text file per opinion; the reason for doing this is that each opinion contains more than one sentence which may express opinion about an important feature. From each opinion sentence, extraction of Opinion Word (OW) is done as shown in fig. 5 [6]. Generally noun defines product feature and adjective defines opinion about that feature. It has been observed that 60-70% opinions are explicit adjectives and 20-30% opinions are explicit nouns, verbs and adverbs [9]. Thus, we require part of speech tagging (POS), to identify nouns, adjectives, verbs, etc. from the opinion sentence [14].

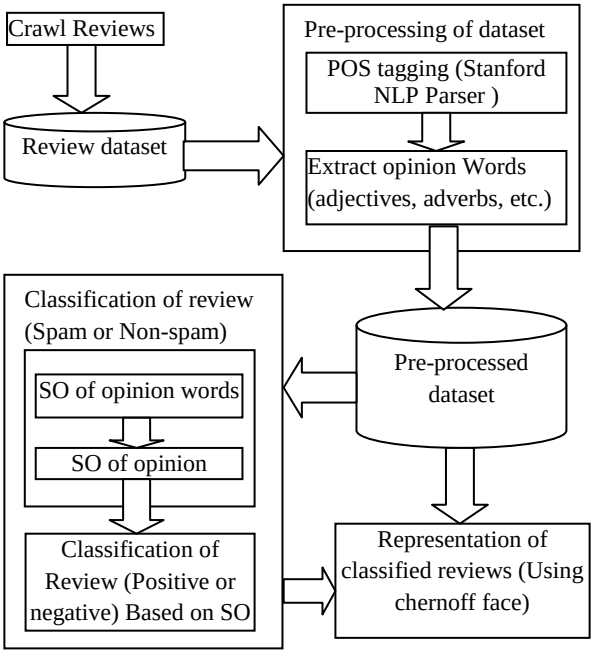


Figure 4: Proposed system architecture for opinion spam detection, removal and visualization

For POS in this paper Stanford NLP parser is used [15].POS tagging will help us in finding product features and also helps for classifying whether the opinion is positive or negative. NLP parser takes one text file as input and parses sentence by sentence and generates dependency tree and structure tree as output shown in fig. 6. Each sentence is saved in the review database along with the POS tag information of each word after stop word removal as shown in fig. 7.

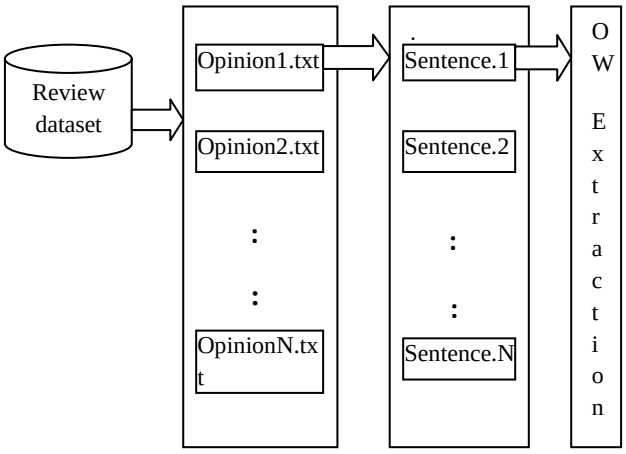


Figure 5: Pre-processing of dataset

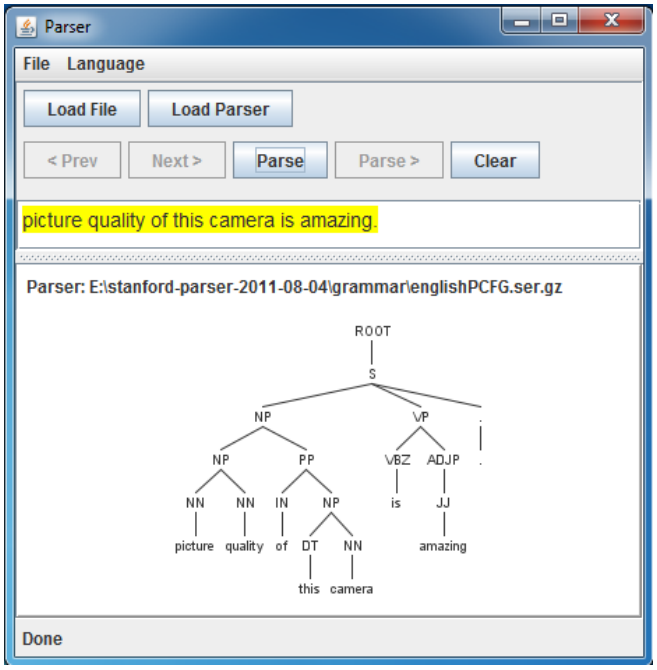
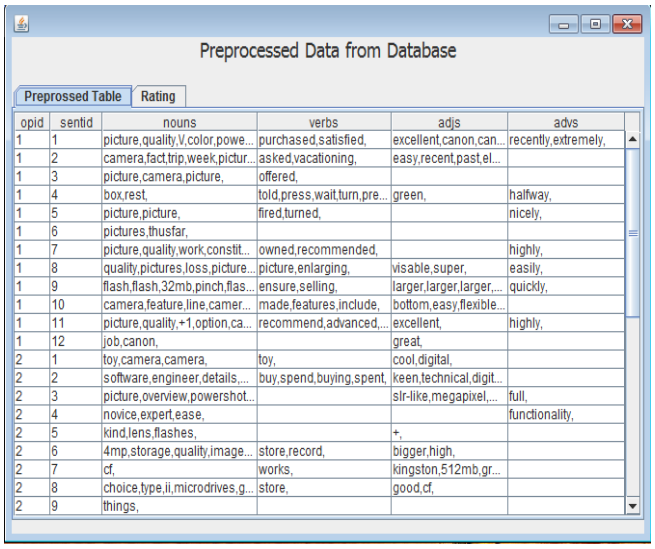


Figure 6: Output of Stanford NLP parser



opid	sentid	nouns	verbs	adjs	advs
1	1	picture,quality,V,color,powe...	purchased,satisfied,	excellent,canon,can...	recently,extremely,
1	2	camera,fact,trip,week,pictur...	asked,vacationing,	easy,recent,past,el...	
1	3	picture,camera,picture,	offered,		
1	4	box,rest,	told,press,wait,turn,pre...	green,	halfway,
1	5	picture,picture,	fired,turned,		nicely,
1	6	pictures,thusfar,			
1	7	picture,quality,work,constit...	owned,recommended,		highly,
1	8	quality,pictures,loss,picture...	picture,enlarging,	visable,super,	easily,
1	9	flash,flash,32mb,pinch,flas...	ensure,selling,	larger,larger,larger,...	quickly,
1	10	camera,feature,line,camer...	made,features,include,	bottom,easy,flexible...	
1	11	picture,quality,+1,option,ca...	recommend,advanced,...	excellent,	highly,
1	12	job,canon,		great,	
2	1	toy,camera,camera,	toy,	cool,digital,	
2	2	software,engineer,details,...	buy,spend,buying,spent,	keen,technical,digit...	
2	3	picture,overview,powershot...		slr-like,megapixel,...	full,
2	4	novice,expert,ease,			functionality,
2	5	kind,lens,flashes,		+,	
2	6	4mp,storage,quality,image...	store,record,	bigger,high,	
2	7	cf,	works,	kingston,512mb,gr...	
2	8	choice,type,ii,microdrives,g...	store,	good,cf,	
2	9	things,			

Figure 7: Pre-processed dataset

3.3 Semantic orientation

In general, adjectives share the same orientation as their synonyms and opposite orientations as their antonyms. We use this idea to predict the orientation of an adjective. To do this, the Synonyms set (synset) of the given adjective and the antonym set are searched. In fig. 9, synset *Fast* consists of two half clusters, one for senses of *fast* and one for senses of *slow*. Each half cluster is headed by *head synset*, in this case *fast* and its antonym *slow* is head synset [9]. If a synonym/antonym has known orientation, then the orientation of the given adjective could be set correspondingly. As the synset of an adjective always contains a sense that links to head synset, the search range is rather large. Given enough seed adjectives with known orientations, we can almost predict the orientations of all the adjective words in the review dataset, i.e. good has 5 bad has -5. Thus, in this research work set of seed adjectives are used, whose orientations are known and then grow this set by searching in the WordNet [16, 17]. Algorithm is given in fig.8 [9] for finding semantic orientation of opinion word.

```
1.procedure OrientationPrediction(adjective_list, seed_list)
2. Begin
3.          do {
4.               size1 = # of words in seed_list;
5.               OrientationSearch(adjective_list, seed_list);
6.                size2 = # of words in seed_list;
7.              }while (size1 ≠ size2);
8. end
1. Procedure OrientationSearch(adjective_list, seed_list)
2.          begin
3.          for each adjective wiin adjective_list
4.                        begin
5.                       if (wi has synonym s in seed_list)
6.                        { wi's orientation= s's orientation;
7.                       add w iwith orientation to seed_list; }
8.                        else if (wi has antonym  a  in seed_list)
9.                       { wi's orientation = opposite orientation
                              of a's orientation;
10.                      add wi with orientation to seed_list; }
11.         end for;
12. end
```

Figure 8: Algorithm for finding semantic orientation of opinion.

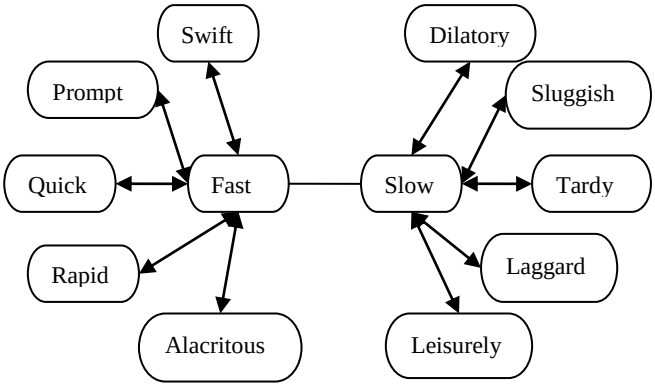


Figure 9: WordNet adjective structure

3.4 Detection of opinion(spam or non-spam)

Each opinion has two main parts: *1. Content* and *2. Rating.* Study shows that rating behaviors are good indicators of spam [3].We can detect possible spam activities based on rating deviations. Two rating behavior which indicates spam is: *1. opinion is highly positive and rating is quite low 2.opinion is negative and rating is high.* For detecting this kind of opinion spam, first classification of opinion whether it is positive or negative is done. From the semantic orientation phase, orientation of each opinion word is known. By using this orientation, classification (positive or negative) of opinion is done. For classification of opinion, semantic orientation of word is given as input. Based on this semantic orientation of opinion is calculated using formula (1) given below.

$$\mathrm{SO}(O_i) = \sum_{k=1}^{n} OW_k \qquad (1)$$

Where SO_i = semantic orientation of i^{th} opinion
O_i = i^{th} opinion
Ow_k = k^{th} opinion word

If opinion's semantic orientation (SO) is a positive value then the opinion is classified as positive opinion and if found negative then opinion is classified as negative.

For detecting opinion spam, comparison of individual rating of opinion with average rating is done. If rating of opinion is varying from average rating then there is possibility that, it can be a spam review. Opinion rating and opinion class (positive or negative) is given as input. Calculation of average rating is done using formula (2) given below.

$$\text{Average Rating} = \left(\sum_{i=1}^{n} OR_i\right)/N \qquad (2)$$

Where, OR_i= i^{th} opinion rating
N=total no. of opinion in dataset

Once the average rating is found, comparison of average opinion rating with individual opinion rating is done. If any deviation is found i.e. opinion has rating 2 and average rating is 4.5 then this can be a spam opinion. This kind of opinion is classified as spam opinion. After classifying opinion as spam, removal of spam opinions is done. Thus after this phase in database genuine opinions will be there.

3.5 Representation of genuine opinion

All genuine opinions are there in dataset, an icon based visual data mining technique "chernoff face" is used for visualizing classified opinions, Dimensions are mapped to the properties of a face as shown in fig. 5.For representing opinion by chernoff face, feature mapping is needed . Let's take one example to understand how to represent data using"chernoff face". Take an example of digital camera. It contains features such as picture quality, size, cost, zooming effect, battery life, weight, memory, etc. Now we can represent each feature of digital camera as one feature of face i.e., picture quality – mouth curvature, cost-eyebrow slope, zoom-eye size, battery life-mouth width shown in fig. 10.

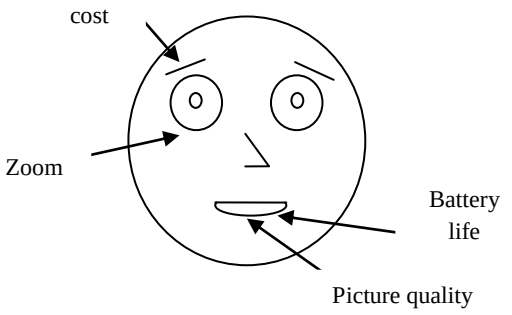


Figure 10: Chernoff face for digital camera

Each chernoff face represents one opinion. Fig. 11 shows five faces which mean five opinions are represented. Increased size of face features represents positive opinion and decreased size represents negative opinion about that feature of camera. In table-2, opinions about digital camera features are analyzed using chernoff face.

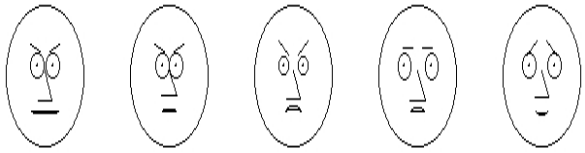


Figure 11: Example of opinion represented using "chernoff face"

Above example was for only few opinions. A large number of faces are then used to represent a data set with one face for each opinion.

Table 2: Opinion about digital camera represented by chernoff face

Opinion No.	Positive	Negative
1	Picture quality, zoom, and battery- life.	Cost.
2	Zoom, battery life	Picture quality, cost
3	Battery life,	Zoom, picture quality, cost
4	Battery life, cost, zoom	Picture quality
5	Battery life, cost, picture quality, zoom	

4. CONCLUSION AND FUTURE WORK

This research work proposes to use "chernoff -faces" to represent genuine opinions after detection and elimination of spam opinions using feature based opinion mining. The use of "chernoff- face" as technique of visualization will help the end user to understand general opinion about specific product in one glance. In the future, we will implement the said "chernoff-face" and analyze the accuracy and precision of results.

5. REFERENCES

[1] Mukherjee, A., Liu, B, Wang, J., Glance, N, Jindal, N. Detecting Group Review Spam.Dept of CS. TechnicalReport, UIC, 2011

[2] Bing Liu, Junhui Wang, Natalie Glance andNitin Jindal. Detecting Group Review Spam . *WWW 2011*, March 28–April 1, 2011, Hyderabad, India.

[3] Lim, E.P., Nguyen, V.A., Jindal, N., Liu, B., and Lauw,H. Detecting product review spammers using rating behavior. CIKM, 2010.

[4] N. Jindal, B. Liu, and E.-P. Lim. Finding unusual reviewpatterns using unexpected rules. CIKM,2010.

[5] Siddu P. Algur, AmitP.Patil, P.S Hiremath and S. Shivashankar.Conceptual level Similarity Measure based Review Spam Detection. IEEE 2010.

[6] Wei Jin, Hung Hay Ho and Rohini K. Srihari. OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction. KDD'09, June 28–July 1, 2009, Paris, France. ACM 2009.

[7] N. Jindal and B. Liu. Opinion spam and analysis. WSDM, 2008.

[8] N. Jindal and B. Liu. Review spam detection.WWW (poster), 2007.

[9] Minqing Hu and Bing Liu. Mining and Summarizing Customer Reviews. KDD'04, August 22–25, 2004, Seattle, Washington, USA. ACM 2004.

[10] Peter D. Turney.Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 417-424.

[11] Web data mining: Web Data Mining Exploring Hyperlinks, Contents, and Usage Data. By Bing liu, First Edition, Dec 2006, Springer.

[12] http://www.amazon.com

[13] Dataset available on: http://www.cs.uic.edu/~liub/FBS /sentiment-analysis.html

[14] POS tagging Penntree bank link: http://www.ling. upenn.edu

[15] NLP parser: http://nlp.stanford.edu

[16] wordnet: http://wordnet.princeton.edu

[17] MIT java api for English wordnet: http://projects.csail .mit.edu/jwi/

[18] A Comparative Study of Visualization Techniques for Datamining:http://www.csse.monash.edu.au/~srini/theses /Redpath_Thesis.pdf

[19] 2D, 3D and High-Dimensional Data and Information Visualization:http://www.iwi.unihannover.de/lv/seminar_ ss05 /bartke/Assets/Paper.pdf