Identification and Analysis of Descriptive Opinion Spam

6th Semester Mini Project

Submitted by

Shubham Sharma (IIT2012134)
Ronish Kalia (IIT2012139)
Shubham Mehrotra (IIT2012156)
Dhruv Kumar (IIT2012171)
Pankaj Wadhwani (IIT2012174)
Arjun Banga (IIT2012183)

Supervisor:

Dr. Ratna Sanyal
IIIT-Allahabad

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD

CERTIFICATE FROM SUPERVISOR

I do hereby recommend that the mini project report prepared under my supervision, titled "**Identification and Analysis of Descriptive Opinion Spam**" be accepted in fulfillment of the requirements of the completion of end-semester of sixth semester of Bachelor of Technology in Information Technology**.**

Date: 28.04.2015 **Guide's name**

Place: Allahabad

_____

Dr. Ratna Sanyal

CANDIDATES' DECLARATION

We hereby declare that the work presented in this project entitled "**Identification and Analysis of Descriptive Opinion Spam**" submit in fulfillment of the 6th Semester of Bachelor of Technology (B. Tech) program, in Information Technology at Indian Institute of Information Technology, Allahabad is an authentic record of our original work carried out under the guidance of Dr. Ratna Sanyal and due acknowledgements have been made in the text of the project to all the other material used. This work was done in full compliance with the requirements and constraints of the prescribed curriculum.

Date:28.04.2015

Place: Allahabad

ABSTRACT

Online reviews provide valuable information about products and services to consumers. However, spammers are joining the community trying to mislead readers by writing fake reviews. Previous attempts for spammer detection used reviewers' behaviors, text similarity, linguistics features and rating patterns. Those studies are able to identify certain types of spammers, e.g., those who post many similar reviews about one target entity. However, in reality, there are other kinds of spammers who can manipulate their behaviors to act just like genuine reviewers, and thus cannot be detected by the available techniques. We aim to minimize the effect of these spam reviews on the overall score of the product by using various techniques and thereby eliminating these reviews or reducing their weightage in the overall score.

Table of Contents

1.1 Introduction

Electronic spamming is the use of electronic messaging systems to send unsolicited messages (spam), especially advertising, as well as sending messages repeatedly on the same site Spamming remains economically viable because advertisers have no operating costs beyond the management of their mailing lists, and it is difficult to hold senders accountable for their mass reviewing. Because the barrier to entry is so low, spammers are numerous, and the volume of unsolicited content has become very high

Online store reviews are an important resource to help people make wise choices for their purchases. Due to this reason, the review system has become a target of spammers who are usually hired or enticed by companies to write fake reviews to promote their products and services, and/or to distract customers from their competitors. Driven by profits, there are more and more spam reviews in major review websites, such as Amazon.in, or Flipkart.com. Spammers are starting to corrupt the online review system and confuse the consumers.

1.2 Objective

In our 5th semester, we were working on a project which dealt with the opinion analysis of customer reviews on E commerce websites. In our output, we had displayed the overall as well as the feature wise rating of the selected product by mining the sentiment of the reviewer.

After analyzing a few reviews manually, we found that they were not coherent with the final output of our study. We figured out that this ambiguity was mainly due to unsolicited reviews which were given by people who were either hired by companies to either benefit the company or used to malign the reputation of their competitors.

As we had individually monitored the features of the product, by finding out which feature has the largest number of spam reviews in it, we would able to indicate which features influence the users in the most impactful way. This data can be used by the company for the marketing of the product.

Previous studies mainly utilize rating as indicator for the detection. However, these studies ignore an important problem that the rating will not necessarily represent the sentiment accurately. In our project, we first incorporate the sentiment analysis techniques into review spam detection.

2.1 *Reasons for Spamming*

- Spammers are usually for profits, so they have connections to stores that would benefit spammers to promote their prominence or defame other stores.
- Spammers are usually hired by low quality stores. Such stores have a stronger motivation to hire spammers to write dishonest reviews. Stores with good reputations and stable customer traffic may not hire spammers at all; since they lose much more if they are caught doing so. Even if good stores really entice spammers to say good things about them, it may not be very harmful. Therefore, we assume that less reliable stores are more likely to be involved in review spamming.
- Harmful spam reviews always deviate from the truth. Therefore, they can be either positive reviews about lousy stores, or negative reviews about good stores.
- Not all reviews deviating from mainstream are spam. People may feel differently or have different experiences about the same service

3. Literature Survey

- In [1], is the first attempt to study of spam detection that gives two methods for spam detection as duplicate detection and spam classification. They consider duplicate review is positive reviews, i.e. spam and others are negative reviews, and they use it for training a model to find out non-duplicate review. But text content is not enough for identification so that they use Naïve Bayes classification to classify spam and non spam review. They find out three types of duplicate positive reviews that used as a spam: (1) duplicates from different user id on the same product, (2) duplicates from the same user id on different products; and (3) duplicates from different user id on different products.

- In [2], they identify three types of spam reviews as untruthful reviews, review on brand only and non-review, then they gave following strategy for spam detection as: First detection of duplicate and near-duplicates using shingle method. The detection of review on brand and non-review is based on machine learning and manual labelled example. Untruthful opinion spam that finds out three types of duplicates.

- In[3], identify eight criteria as Proportion of Positive Singletons (PPS), Concentration of Positive Singletons (CPS), Reactive Positive Singletons (RPS), Review Weighted Rating (RWR), Contribution Weighted Rating (CWR), Truncated Rating (TR), Sentiment Shift (SS), Positive Review Length Difference (PRLD), then find the score matrix with these criteria for all hotels. The aggregation methods are used as Singular value decomposition (SVD) and Unsupervised Hedge algorithm to obtain suspicious review.

- In[4], another work related to spam detection is finding unusual review pattern using Class Association Rules (CAR) that satisfy user given minimum support and minimum

confidence constraints.

In [5], propose three approaches for finding deceptive opinion. First, Genre Identification test has carried out for each review to find out relation frequency distribution of part of speech tags in a text and is depended upon the genre of text. Second, Psycholinguistics Deception detection uses a tool as Linguistics Enquiry and word Count (LIWC) to detect four categories: Linguistics processes to find functional aspects, Psychological Processes to find all social, emotional, cognitive, perceptual and biological processes and anything related to timing and space, Personal Concern considers any references to work, leisure, money, religion, etc. Spoken categories have primarily filler and agreement words. Third, Text categorization approach allows us to model to both content and context with a n-gram features.

- In[6], this is recent work in review spam detection is concerned with a problem of singleton review, i.e. the reviewer written only one review using time series pattern discovery in that they find the correlation between rating and volume of singleton reviews because as the review increases, rating is increases or decreases dramatically. They give a hierarchical framework for robust Singleton review spam detection.

4. Our Proposed Workflow

In continuation of our mini project done in the last semester, we look to analyse spam detected in various customer reviews. Our workflow includes usage of 5 types of spam detection techniques and we eliminate deceptive reviews for a better and more accurate analysis of customer review

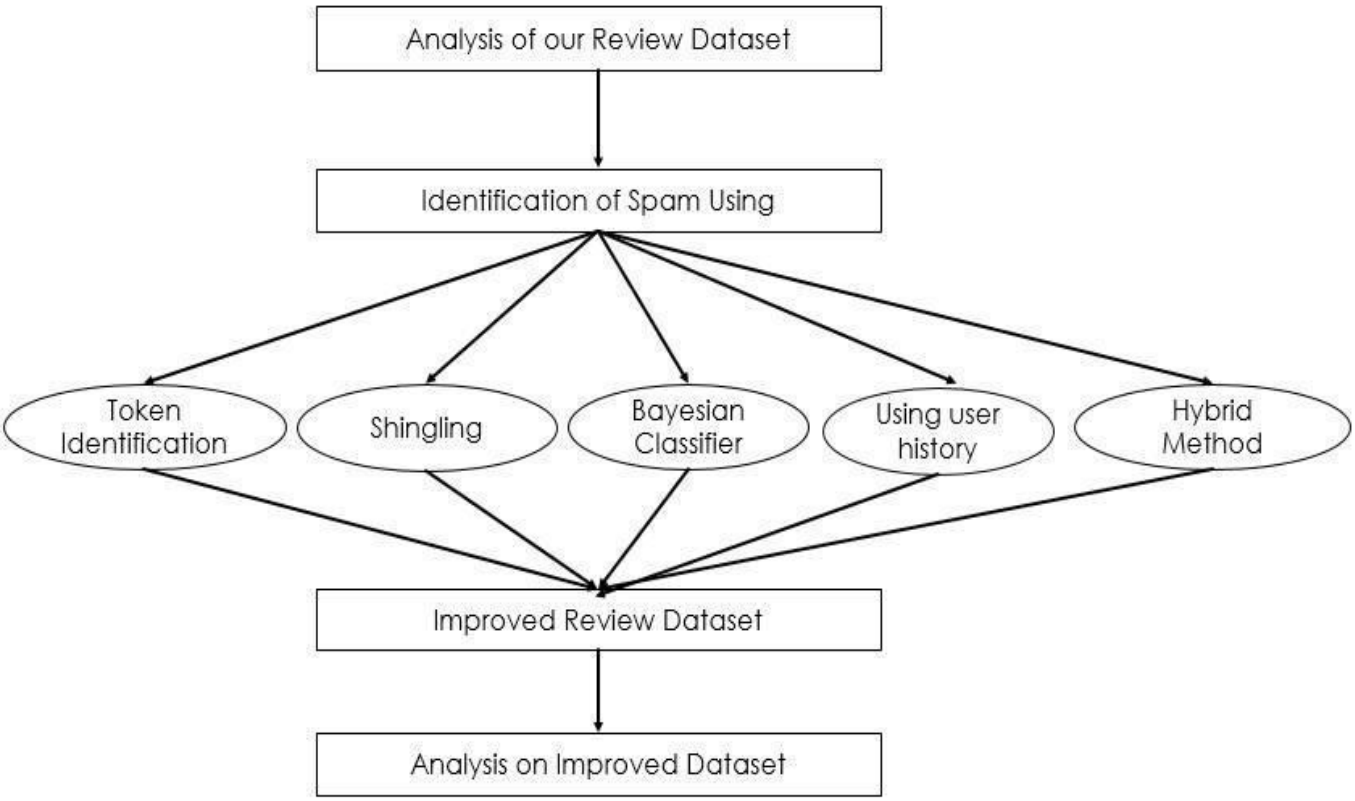

Fig 1: Proposed Workflow

5. Spam Elimination Approaches

5.1 Elimination of Common Spam words:-

This is the first and the most basic technique that we have implemented in our project. This technique eliminates the most trivial type of spams that a user may encounter on an ecommerce website. In this technique we search if the reviews have any vulgar content or are misleading to the user.

The common words that would help us mark the reviews as spam could include swear words. The reviews could also be used to publicize certain set of websites and could also be used to dupe customers. We have used words like share, follow, subscribe, click, etc. to identify these type of reviews. All the reviews that contain these words are marked as spam. Some common spam words have been attached below.

Billion	Incredible Deal
Cash Bonus	Info you requested
Cheap	Information you requested
Collect Child Support	Internet Market
Compare rates	Leave
Compete for your business	Limited time offer
Credit	Make $
Credit Bureaus	Mortgage Rate
Earn $	Obligation
Earn extra cash	Online Marketing
Eliminate Debt	Opportunity
Email marketing	Order Now
Explode your Business	Prices
Extra Income	Promise you
Double your income	No investment

Table1. List of the common spam words

5.2 Shingling Method

This technique detect spam and non-spam reviews based on the product features commented in the reviews.

The two types of spam reviews detected and removed by this technique are:

- Duplicated Review: 2 reviews are considered to be duplicate reviews if their feature match percentage is 100%.
- Near Duplicated Review: 2 reviews are considered to be near duplicate reviews if their feature match percentage is less than 100 percent and greater than or equal to 75 percent.

The other 2 types of non-spam reviews are:

- Partially Related Review: 2 reviews are considered to be partially related reviews if their feature match percentage is less than 80 percent and greater than 0 percent.
- Unique Review: It is the one in which the number of matching feature percentage between two reviews is zero.

*Calculation of feature match percentage will be discussed later.

Steps –

1. Review Pre-processing

Remove the stop words, special characters, punctuations and delimiters occurring in the reviews

2 Feature Extraction

Extract the features from the each review and store it in a database as shown in table below.

Review No.	Extracted Features (fi)
R1	Size colour zoom lcd
R2	ease use zoom weight size
R2	Price lcd lens Photo quality
:	……..
Rm	Color price zoom software

Table2. Extracted Features from a review

3. Create Shingles –

This phase takes as an input the extracted features of the reviews from step 1 and creates its shingles of size w (A contiguous sub sequences contained in Ri is called a shingle). Here any arbitrary review is reduced to a canonical sequence of tokens. (where “canonical” means that any two reviews that differ only in formatting or other information that we chose to ignore, for instance punctuation, formatting commands, capitalization, and so on, will be reduced to the same sequence). So every review Ri is associated with a set of sub sequences of tokens labelled as S (Ri, w). Given a review Ri we can associate to it its w-shingling which is defined as set of all shingles of size w contained in Ri.

So for instance the 2-shingling of the following review is

R1: Sharp photos, size, ease of use;

Pre-processed R1: Sharp photos size ease of use

Extracted features of R1: photos size ease use

2- Shingles of R1 is the following bag (multiset) of all shingles of size 2
{(photos, size), (size, ease), (ease, use)}

Once w is fixed, the resemblance rw (matching feature percentage / 100) of two review documents R1 and R2 is defined as

$$rw(R1, R2) = \frac{|S(R1, w) \cap S(R2, w)|}{|S(R1, w) \cup S(R2, w)|}$$

The resemblance is some number between 0 and 1.

4. Spam Detection:

For each review Ri in the

Parse Ri and generate a unique md5 hash code for each of the generated
shingles. For all the remaining reviews Ri+1to Rm

{

Parse each of Ri+1 to Rm reviews and generate a unique md5 hash code for each of the generated shingles.

Compute the resemblance ratio of the hash code values of the shingles between reviews Ri and Ri+1to Rm using the following formulae

$$rw(Ri, Ri+1 \text{ to } Rm) = \frac{|S(Ri, w) \cap S(Ri+1 \text{ to } Rm, w)|}{|S(Ri, w) \cup S(Ri+1 \text{ to } Rm, w)|}$$

Classify the reviews as spam or non-spam based on resemblance ratio percentage.

Review pairs with resemblance (similarity score) of 100% are chosen as duplicates (i.e. two reviews which have similar features being commented are called duplicates). Similarity score of 75 % to 99% between the reviews are chosen as near duplicates (i.e. two reviews which have almost similar features being commented are called near duplicates). Similarity score of less than 75 % between the reviews are chosen to be partially related.

Algorithm –

{

Parse Ri to create tokens (shingles) of size w i.e. S (Ri, w) and generate a unique md5 hash code for each of the generated shingles.

For all the remaining reviews Ri+1to Rm

{

Parse each of Ri+1 to Rm reviews to create tokens (shingles) of size w i.e. S (Ri+1, w), S (Ri+2, w)….. S (Rm, w) and generate a unique md5 hash code for each of the generated shingles.

Compute the resemblance ratio of the hash code values of the shingles between reviews Ri and Ri+1to Rm using the following formulae

$$rw(Ri, Ri+1 \text{ to } Rm) = \frac{|S(Ri, w) \cap S(Ri+1 \text{ to } Rm, w)|}{|S(Ri, w) \cup S(Ri+1 \text{ to } Rm, w)|}$$

Classify the reviews as spam or non spam based on resemblance ratio percentage

}

}

Example:

Consider the following two reviews

R1 = Easy to learn functions! Size!

R2 = small size, great functions

Pre-processed R1 = Easy to learn functions Size

R2 = small size great functions

Extracted Features of R1 & R2

R1 = functions size

R2 = size functions

1-shingling of the above features extracted from the two reviews is the bag of all shingles of size 1 contained in R1 and R2 written as

S (R1, 1) = {(functions), (size)}

S (R2, 1) = {(size), (functions)}

$$rw(R1, R2) = \frac{|S(R1, 1) \cap S(R2, 1)|}{|S(R1, 1) \cup S(R1, 1)|}$$

$$rw(R1, R2) = \frac{|2|}{|2|} = \frac{2}{2} = 1$$

{100 % similarity}

R1 resembles 100% with R2 for shingle size 1, and hence are considered as duplicate spam reviews

5.3 Bayesian Classifier

Let's suppose the suspected message contains the word “subscribe". Most people who will read a review would know that this review is likely to be spam. The spam detection software, however, does not "know" such facts; all it can do is compute probabilities.

The formula used by the software to determine that is derived from Bayes' theorem

$$\Pr(S|W) = \frac{\Pr(W|S) \cdot \Pr(S)}{\Pr(W|S) \cdot \Pr(S) + \Pr(W|H) \cdot \Pr(H)}$$

where:

- $\Pr(S|W)$ is the probability that a message is a spam, knowing that the word "replica" is in it;
- $\Pr(S)$ is the overall probability that any given message is spam;
- $\Pr(W|S)$ is the probability that the word "replica" appears in spam messages;
- $\Pr(H)$ is the overall probability that any given message is not spam (is "ham");
- $\Pr(W|H)$ is the probability that the word "replica" appears in ham messages.

The spamicity of a word

Recent statistics show that the current probability of any review being spam is 80%, at the very least:

$$\Pr(S) = 0.8; \Pr(H) = 0.2$$

However, most bayesian spam detection software makes the assumption that there is no *a priori* reason for any review to be spam rather than ham, and considers both cases to have equal probabilities of 50%

$$\Pr(S) = 0.5; \Pr(H) = 0.5$$

The filters that use this hypothesis are said to be "not biased", meaning that they have no prejudice regarding the current review. This assumption permits simplifying the general formula to:

$$\Pr(S|W) = \frac{\Pr(W|S)}{\Pr(W|S) + \Pr(W|H)}$$

This is functionally equivalent to asking, "what percentage of occurrences of the word "subscribe" appear in spam reviews?"

This quantity is called "spamicity" (or "spaminess") of the word "subscribe", and can be computed. The number $\Pr(W|S)$ used in this formula is approximated to the frequency of messages containing "replica" in the messages identified as spam during the learning phase. Similarly, $\Pr(W|H)$ is approximated to the frequency of reviews containing "subscribe" in the reviews identified as ham during the learning phase. For these approximations to make sense, the set of learned reviews needs to be big and representative enough. It is also advisable that the learned set of reviews conforms to the 50% hypothesis about repartition between spam and ham, i.e. that the datasets of spam and ham are of same size.[9]

Of course, determining whether a message is spam or ham based only on the presence of the word "subscribe" is error-prone, which is why bayesian spam software tries to consider several words and combine their spamicities to determine a review's overall probability of being spam.

Combining individual probabilities

Most bayesian spam filtering algorithms are based on formulas that are strictly valid (from a probabilistic standpoint) only if the words present in the review are independent events. This condition is not generally satisfied (for example, in natural languages like English the probability of finding an adjective is affected by the probability of having a noun), but it is a useful idealization, especially since the statistical correlations between individual words are usually not known. On this basis, one can derive the following formula from Bayes' theorem

$$p = \frac{p_1 p_2 \cdots p_N}{p_1 p_2 \cdots p_N + (1-p_1)(1-p_2)\cdots(1-p_N)}$$

where:

- p is the probability that the suspect message is spam;
- p_1 is the probability $p(S|W_1)$ that it is a spam knowing it contains a first word (for example "replica");
- p_2 is the probability $p(S|W_2)$ that it is a spam knowing it contains a second word (for example "watches");
- etc...
- p_N is the probability $p(S|W_N)$ that it is a spam knowing it contains an *N*th word (for example "home").

The result *p* is typically compared to a given threshold to decide whether the message is spam or not. If *p* is lower than the threshold, the message is considered as likely ham, otherwise it is considered as likely spam.

5.4 Using User Review History

Previous attempts for spammer detection used reviewers' behaviors, text similarity, linguistics features and rating patterns. Those studies are able to identify certain types of spammers, e.g., those who post many similar reviews about one target entity. However, in reality, there are other kinds of spammers who can manipulate their behaviors to act just like genuine reviewers, and thus cannot be detected by the available techniques.

The main reason is that spammers can easily disguise themselves. It is thus hard for a human user to recognize them, while for Web and email spam, one can tell spam without much difficulty.

- There is *no ground truth* whether a review is faked or not. By reading the review text alone, we usually do not have enough clues to tell spam from non-spam.
- Spammers' behaviors may be hard to capture. For example, in order to successfully mislead customers, spammers can make their writing styles and review habits look very similar to those of genuine reviewers.
- Spammers can also write good and honest reviews, because they could be real customers.

So we will use a model wherein we will use various factors such as trustiness of the user and honesty of the review. This technique will require a training data set and will become more and more efficient over the period of time.

Our contribution to this method has been that we are considering the fact that multiple reviews on a particular product by the same user be considered suspicious and we are reducing the honesty/ trustiness of the reviewer exponentially for each review on the same product. We are also emphasizing on the reliability of the product i.e. brand instead.

$$T(User) = K/(1 + e^{-KH_{r \sim \Sigma}})$$

$$T(User) = T(User) * (1 / (1 + e^{x} - 1)) \quad R(P) = 2 / (1 + e^{-Q}) - 1$$

5.5 Hybrid Approach

In this model, we have integrated the user history into the Bayesian method. We have taken advantage of the fact that two spamicity scores provided by the above two methods, i.e. the Bayesian method and Graphical Review method when used together could give us a better result for whether a certain review is a spam or not.

We are taking review honesty, user trustiness and product reliability as separate fields in the Bayesian Method and then computing a new spamicity score while considering all the previous features as well.

New Spamicity Score = RHS * UTS * PRS * BS

RHS: Review honesty score

UTS: User Trustiness Score

PRS: Product Reliability Score

BS: Bayesian Score

This method would have the benefits of both the methods thus giving us a better and improved score.

6. Results and Analysis

We had the initial dataset on which we applied various spam detection techniques. Firstly, we have applied Bayesian Technique, then graphical method of spam detection (which uses users review history) and finally the hybrid approach which is also the improved version of Bayesian combined with graphical method.

Different types of spams are removed by different techniques which have resulted in improving the accuracy and hence provided the genuine rating of the attributes of a product.

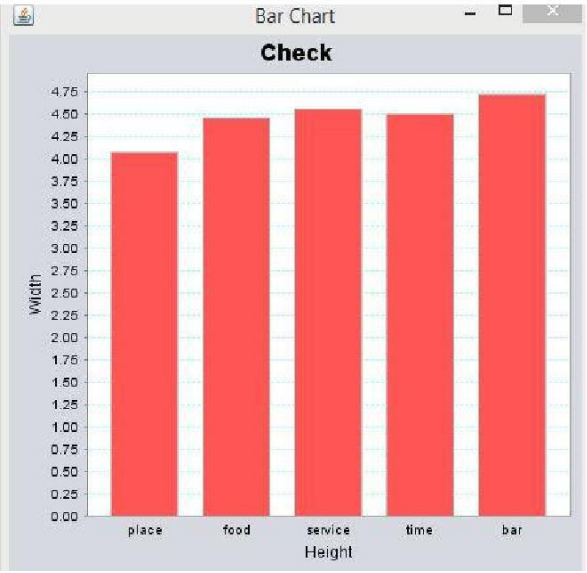

Fig2. Before applying shingling algorithm

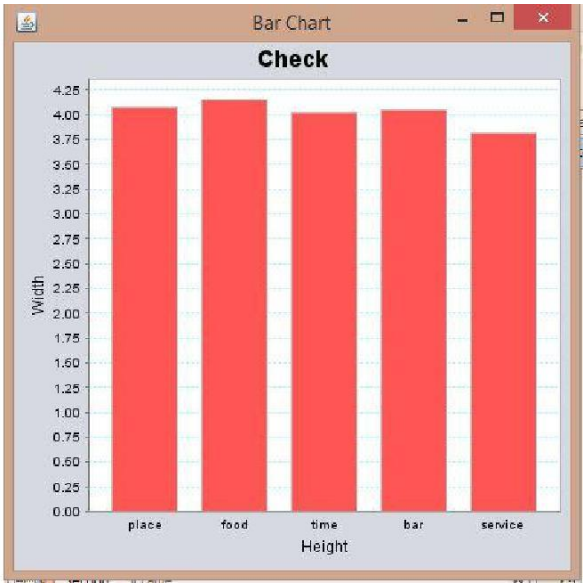

Fig3. After applying shingling algorithm

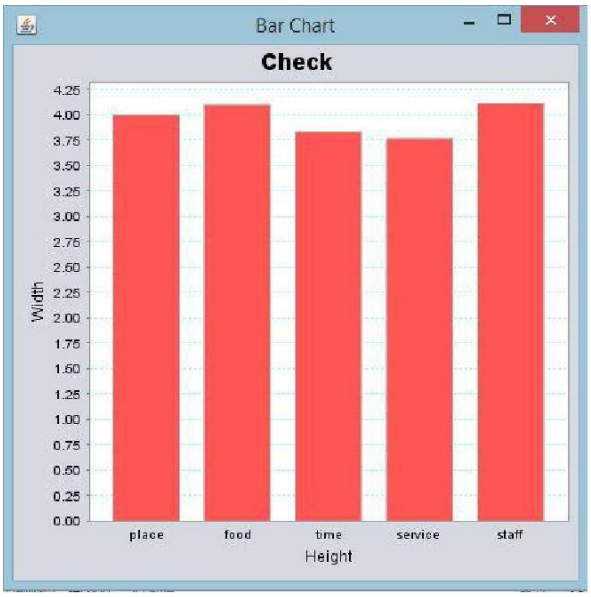

Fig4. Initial dataset on which various techniques are performed

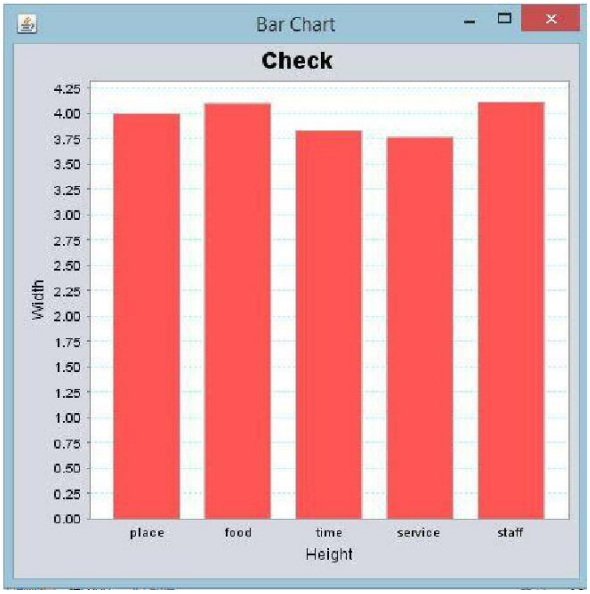

Fig5. Results after applying Bayesian technique

```
place 1270
food 1011
time 606
service 461
staff 238
4.003772520602604
4.083334701342202
3.828658241871631
3.7528097427770546
4.121717157044677
```

Fig 6. Feature rating of initial dataset

```
Reading POS tagger mod
place 920
food 753
time 430
service 335
staff 182
3.9964395543038336
4.0977542119817
3.826995031864354
3.7623819966426296
```

Fig 7. Rating after applying Bayesian

```
place 1066
food 847
time 496
service 373
staff 200
3.98424763084744 57
4.104556778296287
3.8133807542823743
3.785850123916626
4.14418946080054
```

Fig8. Rating after applying graphical method (user review history)

```
time 496
service 373
staff 200
3.855015661485619
4.0043788778981 98
3.678847329924459
3.6660329766904005
4.039758142966232
```

Fig9. Rating after applying hybrid approach

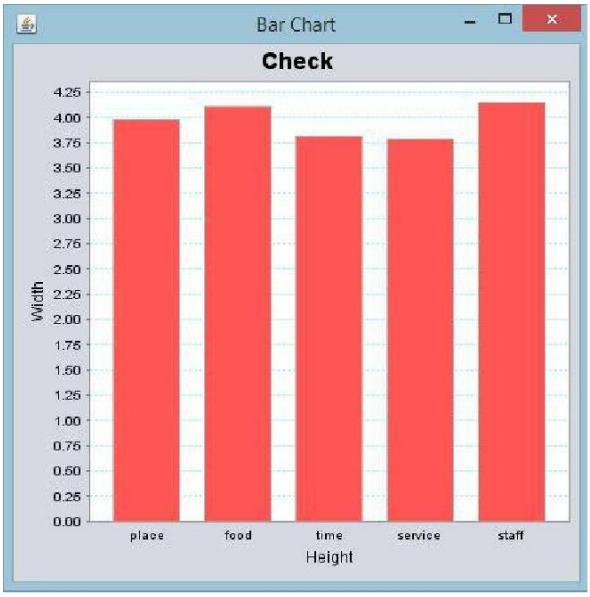

Fig 10. Results after graphical approach

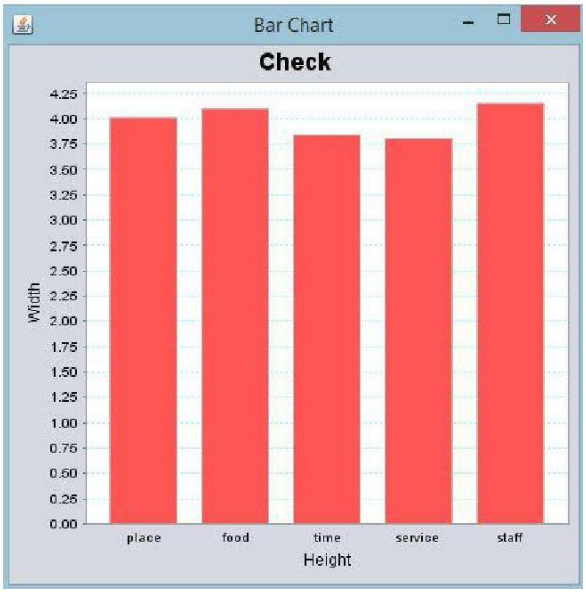

Fig11. Results after the hybrid method

7 .References

*1+Nitin Jindal, Bing Liu ,"Review Spam Detection" , *ACM Proceedings of the 16th international conference on World Wide Web*, pp-1189-1190,2007.

*2+ Nitin Jindal, Bing Liu," Opinion Spam and Analysis",*ACM Proceedings of the international conference on Web search and web data mining*, pp.219-229, 2008.

*3+ Guangyu Wu, Derek Greene, Pádraig Cunningham ,"Merging multiple criteria to identify suspicious reviews", *Proceedings of the fourth ACM conference on Recommender systems*, pp.241-244,2010.

[4] Nitin Jindal, Bing Liu, Ee-Peng Lim "Finding unusual review pattern using unexpected rules", *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp.1549-1552,2010.

*5+ Myle Ott, Yejin Choi, Claire Cardie, Jeffrey T. Hancock, "Finding deceptive opinion spam by any stretch of imagination" , *ACM Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* , pp.309-319,2011.

*6+ Sihong Xie, Guan Wang, Shuyang Lin, Philip S. Yu "Review spam detection via time series pattern discovery", *ACM Proceedings of the 21st international conference companion on World Wide Web*,pp.635-636,2012.

[7] Raymond Y. K. Lau, S. Y. Liao, Ron Chi-Wai Kwok, Kaiquan Xu, Yunqing Xia, Yuefeng Li," Text mining and probabilistic modeling for online review spam detection" *ACM Transactions on Management Information Systems (TMIS)* , Volume 2 Issue 4,Article 25,2011.

8. Suggestion by Board Members