Review Spamicity approach by Resemblance Measure

Rajshree
Department of Computer Science, Government College (Autonomous), Kalaburgi

***Abstract*—The ubiquity of web2.0 makes the web an invaluable source of information. For instance, product reviews composed collaboratively by many independent internet reviewers can help consumers make purchase decisions and enable enterprises to improve their business. In this work, an attempt is made to compare and detect whether a review is spam or non-spam review from different websites, in order to have a mechanism for proper decision making or for marketing intelligence.**

Keywords - Reviews, Feature extraction, Opinion mining, Spam.

I.INTRODUCTION

Understanding the content of the review from the reviewers relating to particular product, is the "key concept" being expressed[4]. Locating the topic, main idea, and supporting details helps one to understand the point(s) or reviews. Identifying the relationship between the reviewers and reviews will increase our comprehension [11]. The web contains a wealth of opinions about products, politicians, and more, which are expressed in newsgroup, posts, review sites, and elsewhere. As a result, the problem of opinion mining has seen increasing attention over the last decade. Product reviews on web sites such as amazon.com, cnet.com and epinion.com and elsewhere often associate meta-data with each review indicating how positive (or negative) it is using a 5-star scale, and also rank products by how they fare in the reviews at the site[6].It is now a common practice for E-Commerce web sites to enable their customers to write reviews of products that they have purchased. The reviews are then used by potential customers to find opinions of existing users before purchasing the products[13]. They are also used by manufacturers to identify problems in their products and/or to find competitive intelligence information about their competitor [2][3]. The number of customer reviews that a product receives is growing at a very fast rate. 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, which is crucial for all opinion based applications, although web spam and email spam have been investigated extensively. Different websites provide different formats for writing the reviews. There are three different types of review formats available on the web. Format (1) - Pros and Cons: The reviewer is to describe Pros and Cons separately.Cnet.com uses this format. Format (2) - Pros, Cons and detailed review: The reviewer is to describe Pros and Cons separately and write a detailed review, Epinions.com uses this format. Format (3) - free format: The reviewer can write freely, i.e., no separation of Pros and Cons, Amazon.com uses this format. In this work, we aim to summarize customer reviews of a product from various websites like Cnet.com and Epinion.com, etc., for the same product.

II. RELATED WORK

In [15] it gives a web mining taxonomy but restricted to web content and web usage mining and gives a survey on web usage mining. It divides the web content mining into the agent based approach and the database approach. Most relevant work in review mining is that of (Hu and Liu, 2004) [1]. At present Opinion Mining has become a vital research subject in the field of product reviews. [4]Although mining opinions (positive and negative) from reviews became a popular research topic in recent years [1,5] there is still no reported study on review spam. A taxonomy of Web spam is given in [5].Few researchers have studied this problem [e.g., 1, 5, 6]. Review spam is very different. Adding irrelevant words has little effect. Instead, spammers write undeserving positive reviews to promote some objects and/or malicious negative reviews to damage the reputation of some other objects. These false opinion spam reviews are very hard to detect. Another related research is email spam [7, 8], which is also quite different from review spam. Email spam usually refers to unsolicited commercial advertisements. Although exist, advertisements in reviews are not as frequent as in emails. Recent studies

on spam also extended to recommender systems [9]. Although the objectives of spam on recommender systems are similar to review spam, their basic ideas are different. In recommender systems, a spammer injects some attack profiles to the system in order to get some products more (or less) frequently recommended. A profile is a set of ratings (e.g., 1-5) for a series of products. The spammer usually does not see other users rating profiles and thus has to make guesses. In the context of product reviews, a reviewer sees all reviews for every product. Rating is only part of a review and another main part is the review text [10] studies the utility of reviews using natural language features. Spam is a much broader concept involving all types of objectionable activities.

III.PROPOSED METHODOLOGY

We propose a novel and effective technique to extract the customer reviews from multiple web sites (Cnet.com and Epinion.com, ect) for the same product and compare and detect spam and non-spam reviews in it based on the product features similarity [18] that have been commented in the reviews. The idea is based on mining review features in customer reviews from different websites and finding the spamicity between the reviews based on conceptual level similarity measure. The spamicity and non-spamicity is based on two types of spam and non-spam reviews. The two types of spam reviews are: (i) Duplicated Review: If the set of features (concepts) corresponding to the two reviews are exactly identical (i.e. 100% duplication of features), then the two reviews are said to be duplicated reviews. (ii) Near Duplicated Review: The number of matching features corresponding to the two reviews are less than 100% and between certain specified threshold (i.e. not an exact copy, but almost similar content), then the two reviews are said to be near duplicated reviews. Two types of non- spam reviews are: (i) Partially Related Review: The number of matching features corresponding to the two reviews is less than the specified threshold, then the two reviews are said to be partially related reviews. (i.e. both the reviews resemble in very few common features in its opinion). (ii) Unique Review: It is the one in which the number of matching features between the two reviews is zero, i.e. there is no duplication of the features between the two reviews Thus the two reviews are totally unrelated or unique in their opinion.

The various components of the proposed method include:

1 Review Extraction and Store
2 Feature Mining.
3 Similarity Measure

3.1 Review Extraction and Store

Input to the review extraction and data store component is the webpage containing the reviews and output is the extracted reviews that are stored in the raw review database.This component is common to all further components in the system model in providing the raw reviews for detecting spam or non-spam reviews in it extracted from multiple websites such as Cnet.com, Epinion.com, etc.

3.2 Feature Mining.

It takes raw reviews as input stored in the raw review database extracted from multiple websites and mines the product features using the existing methods [17] and constructs the feature matrix and ranks them using histogram for further processing. It has the subcomponents namely feature extraction, feature matrix construction and feature ranking.

3.3 Similarity measure

In this component, we compare and detect whether reviews from multiple websites are spam or non-spam reviews using conceptual level similarity measure. This component accepts the feature matrix as the input and finds percentage of matching of features from one website to another to detect them as spam or non-spam. Conceptual level is an ontology based similarity [22], which takes care of conceptually similar words by mapping words to concepts. The concepts are the features extracted from the reviews. The constructed feature matrix, from both the websites Cnet.com and Epinion.com are read as the input. Let RC={RC1,RC2,RC3….RCM}be the reviews with its features extracted from Cnet.com and RE={RE1,RE2,RE3….REN} be the reviews with its features extracted from Epinion.com. In order to detect the spam and non-spam using similarity measure from two websites, the two feature matrix are to be compared to find the matching number of features or its equivalent synonyms between the reviews of both matrix. Concepts pertaining to the two reviews could be exactly identical (i'e duplicates) nearly duplicates, partially related or unique as indicated in the Table 1.

Table 1: Conceptual level similarity cases

	Similarity Case	Examples / Documents Maximum number of concepts =20 Concepts(f_1,f_2,f_3,f_4,f_5……f_{20})	Similarity Value
Spam	Duplicate	Rci{1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1} Rej{1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1}	17
		Rci{1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1} Rej{1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1}	16
	Near Duplicates	Rci{1,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1} Rej{0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1}	14
		Rci{1,0,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,0,1,1} Rej{1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,0}	12
Non Spam	Partially related	Rci{1,0,1,0,1,1,1,0,1,1,0,1,1,1,0,1,1,0,1,1} Rej{1,0,1,0,1,0,1,1,1,1,1,1,1,0,1,1,0,1,1,0}	8
		Rci{1,1,1,1,1,1,1,0,1,1,0,1,1,1,0,0,1,0,0,1} Rej{1,0,1,0,1,0,1,1,1,1,1,1,1,0,0,1,1,0,1,1,0}	7
	Unique reviews	Rci{1,0,1,1,0,1,1,0,0,1,0,0,1,1,0,0,1,1,0,0,1} Rej{0,1,0,0,1,0,0,1,1,0,1,1,0,0,1,1,0,1,1,0}	0
		Rci{1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1} Rej{0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0}	0

Conceptual level similarity measure between two review documents RCi and REk is defined as follows.
sim (RCi, REk) = NC – DH (RCi, REk)
where NC = Total number of feature (concepts) in each review RCi. DH(RCi, REk) = Hamming distance between review RCi and REk. This similarities measured is used to classify the reviews as spam and non-spam based on a threshold value T to be data mined empirically. The classification rule is given by:
(i) (i) If sim (RCi, REk) is in [T, NC], then RCi and REk are spam.
(ii) Further for spam reviews RCi and REk, if sim (RCi, REk) = NC, then the reviews RCi and REk are duplicates and near duplicates.
(iii) Similarly, for non-spam review RCi and REk, if sim(RCi, REk)= 0, then the review RCi and REk, are unique, otherwise partially related. If the set of concepts corresponding to the two reviews are exactly identical, then the two reviews are said to be duplicated at the conceptual level and if the number of matching concepts between the two reviews is less than 'm' the maximum number of the concepts and is greater than or equal to ten (the threshold is fixed as ten based on experimental observations) then the two reviews are said to be near duplicated at conceptual level.If the set of matching features between the two reviews is less than ten then the two reviews are said to be partially related at conceptual level i.e.subset of the concepts of one review document matches with the subset of the concepts of the other review document. If the threshold between the two reviews is zero, then both the reviews are said to be unique reviews. Here the first review from FM1 is compared with all the reviews from FM2 in order to detect the spam and non-spam review(i'e RC1could be compared with all RE's and if the number of concepts matches as mentioned in the Table.1 for spam, then they will be marked as spam and counted for spam review. The comparison may be either RC's with RE's or RE's with RC's).
However, the proposed method will not be able to find the source of spam i'e is it from Cnet.com or Epinion.com websites. The comparison of matching concepts is demonstrated in Table 2.

Table 2 : Comparison of Matching concepts

Feature Matrix of RC_1 with all RE's	*screen*	*camera*	*battery*	*speed*	..	*quality/apps*	*Total*	*Compared features*	*% of feature*
Review No	f_1	f_2	f_3	f_4	..	F_n	$\sum f_n$		
RC_1	0	1	0	1	..	1	4	4	100
RE_1	0	1	0	1	...	0	6	2	50
RE_2	1	1	0	1	...	1	4	1	25
:	:	:	:	:	:	:	:		
RE_N	0	0	0	1	...	0	3	2	50

Algorithm 1. Detects spam and non-spam reviews using conceptual level similarity measure.

```
Conceptual_level_similarity(FM1,FM2)
// FM1[M:m] is a feature matrix for RC
// FM2[N:n] is a feature matrix for RE
// M is the number of reviews in FM1
// N is the number of reviews in FM2
// m is the number of features (concepts)
     for FM1
// n is the number of features (concepts)
     for FM2
{
For each review RCi in the feature matrix
FM1
{
  For each feature fci of RCi
  {
Compute sim(RCi,REk) for all REk  i<k<=M
Apply classification rule:
Mark the spam review, which are not considered for
future computation of similarity measure.
```

Count the number of spam reviews (marked) and non-spam reviews.

```
        }
    }
}
```

4. EXPERIMENTAL RESULTS

The experimental results are shown in Table 3. Experiments are carried by taking reviews from both websites(Cnet.com&Epinion.com)and detected as duplicates, near duplicates reviews as spam reviews and partially related and unique reviews as non-spam reviews. We considered 10,0000 customer reviews on Samsung galaxy mobile phone, in which every review is compared with all the other reviews in the dataset using the conceptual level similarity measure to find spam and the non-spam reviews. The experimental results shown in Table below depicts the summary of detected spam and non-spam reviews for both the pros and cons reviews on mobile product using the proposed technique. A total of 10000 reviews (5500 reviews from Cnet.com and 4500 reviews from Epinion.com websites) on mobile product are been assessed using conceptual level similarity measure for detecting the spam and non-spam reviews in it. The experimental results show that, there are larger numbers of near duplicate spam reviews detected using the conceptual level similarity measure. These near duplicate spam reviews are untrustworthy as it does not provide genuine opinion about the product and it misleads potential buyer in making his buying decision. From the results we also conclude that there are large numbers of reviews belonging to non-spam category i.e. partially related and unique reviews.

Table 3: Summary of the experimental results

Proposed Techniques		Conceptual level similarity measure	
Types of Reviews		Pros	Cons
Total Reviews		5500	4500
Spam Reviews	Duplicates	0	1200
	Near duplicates	2450	1250
Non Spam Reviews	Partially related	2200	1700
	Unique reviews	850	350

These reviews do not influence the buying decision significantly and hence such reviews could be considered trustworthy as they provide a genuine opinion on some or the other unique feature of the product and are often unbiased.

5. CONCLUSION

In this paper, we proposed a novel and effective technique for detecting the trustworthiness of customer reviews for a particular product (Samsung galaxy mobile). The idea is based on mining review features in customer reviews from various websites Cnet.com, Epinion.com etc., and for finding the spamicity between the reviews based on conceptual level similarity measure. This is performed in three steps (1) Feature extraction (2) Feature mining and (3) Similarity Measure. Product features are considered for spam detection as through observation, it is noted that features describe the product best. This work can be summarized as duplicate and near duplicate reviews being categorized as spam reviews and partially related and unique reviews being categorized as non-spam reviews. Experimental results demonstrate the effectiveness of the proposed technique detecting spam and non- spam reviews. The efficiency of the task of web based customer review spam detection can be enhanced by identifying and eliminating duplicate and near duplicate spam reviews, thereby providing a summary of the trusted reviews for customers to make buying decisions. Detecting spam reviews in case of review format (1) and format (3) and using the semantic analysis towards the review spam detection gives the scope for future work.

REFERENCE

[1] Hu and B. Liu. 2004. Mining and Summarizing Customer Reviews. In KDD, pages 168–177, Seattle, WA.

[2] Siddu P.Algur, Amit P.Patil, P.S Hiremath, S.Shivashankar “Spam Detection of Customer Reviews from Web Pages” in IT and Business Intelligence.

[3] Nitin Jindal and Bing Liu “Analyzing and Detecting Review Spam” in ICDM 2007 IEEE.

[4] Lijia Liu. Yu Wang, “A method for sorting out the spam from Chinese product.” 2012 IEEE.

[5] K. Dave, S. Lawrence & D. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews.WWW’2003.

[6] Siddu P. Algur. , and Jyoti G.Biradar., “Rating Consistency and Review Content based Multiple stores Review Spam Detection” In IEEE InternationalConference on Information Processing”. ISBN 978-1-4673-7758-4/15, 2015

[7] I. Fette, N. Sadeh-Koniecpol, A. Tomasic. Learning to Detect Phishing Emails. WWW'2007.
[8] M. Sahami, S. Dumais, D. Heckerman and E.Horvitz. A Bayesian Approach to Filtering Junk {E}-Mail. AAAI Tech. Report WS-98-05, 1998.
[9] A. Ntoulas, M. Najork, M. Manasse & D.Fetterly. Detecting Spam Web Pages through Content Analysis.WWW'2006.
[10] Z. Zhang & B. Varadarajan, Utility scoring of product reviews, CIKM'2006.54524.
[11] Nitin Jindal and Bing Liu "Opinion Spam and Analysis"WSDM'08, February 11-12, 2008, Palo Alto, California, USA.
[12] Hu and B. Liu. 2004. Mining and Summarizing Customer Reviews. In KDD, pages 168–177, Seattle, WA.
[13] Siddu P.Algur, Amit P.Patil, P.S Hiremath, S.Shivashankar "Spam Detection of Customer Reviews from Web Pages" in IT and Business Intelligence.
[14] Nitin Jindal and Bing Liu "Analyzing and Detecting Review Spam" in ICDM 2007 IEEE.
[15] Lijia Liu. Yu Wang, "A method for sorting out the spam from Chinese product." 2012 IEEE.
[16] K. Dave, S. Lawrence & D. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews.WWW'2003.
[17] Z. Gyongyi and H. Garcia-Molina. Web Spam Taxonomy. Tech. Report, Stanford University, 2004.1sted.Chichester,U.K.:Wiley,1993,ch.2,pp.45-47.
[18] Siddu P. Algur, Amit P.Patil, P.S Hiremath, S. Shivashankar "Conceptual level Similarity Measure based Review Spam Detection" 2010 IEEE.
[19] Popescu, Ana-Maria, and O Etzioni. "Extracting product features and opinions from reviews."HLTEMNLP. 2005.
[20] Hiremath, P.S Benchalli, S.S., Algur Siddu P., and Renuka Udapudi. "Mining Data Regions from Web Pages" COMAD-2005.
[21] Kupiec, J, J Pedersen, and F Chen. "A Trainable Document Summerizer." SIGIR - 1995. 1995.
[22] H.L.Shashirekha, S.Murali, and P. N Agabhushan "Ontology based similarity measure for text documents" International Conference on Signal and Image processing (ICSIP-2009).
[23] Zhongwu Zhai, Bing Liu, Hua Xuand Peifa Jia. "Grouping Product Features Using Semi-Supervised Learning with Soft-Constraints".
[24] P.S.Hiremath, Siddu P.Algur, S.Shivashankar "Cluster Analysis of customer Reviews Extracted from Web Pages".JACSM-2010.
[25] P.S.Hiremath, Siddu P.Algur, S.Shivashankar. "Quality Assessment of Customer Reviews Extracted From Web Pages : A Review Clustering Approach.
[26] P.S.Hiremath, Siddu P. Algur, S.Shivashankar "Web based Quality Assessment of Customer Reviews using Quartile Measure".
[27] Benchalli, S.S, P.S Hiremath, Siddu Algur, and Renuka Udapudi. "Mining Data Regions from Web Pages."COMAD2005b. 2005 DEC.