International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:4, No:2, 2010

Automatic Extraction of Features and Opinion-Oriented Sentences from Customer Reviews

Khairullah Khan, Baharum B. Baharudin, Aurangzeb Khan, and Fazal_e_Malik

***Abstract*—**Opinion extraction about products from customer reviews is becoming an interesting area of research. Customer reviews about products are nowadays available from blogs and review sites. Also tools are being developed for extraction of opinion from these reviews to help the user as well merchants to track the most suitable choice of product. Therefore efficient method and techniques are needed to extract opinions from review and blogs. As reviews of products mostly contains discussion about the features, functions and services, therefore, efficient techniques are required to extract user comments about the desired features, functions and services. In this paper we have proposed a novel idea to find features of product from user review in an efficient way. Our focus in this paper is to get the features and opinion-oriented words about products from text through auxiliary verbs (AV) {is, was, are, were, has, have, had}. From the results of our experiments we found that 82% of features and 85% of opinion-oriented sentences include AVs. Thus these AVs are good indicators of features and opinion orientation in customer reviews.

***Keywords*—**Classification, Customer Reviews, Helping Verbs, Opinion Mining.

I. Introduction

OPINION is a private state of a person thinking about something [1]. When we state about something we express our thoughts about that particular thing on the basis of our observations, knowledge and experience. Such statement about things is useful for those who are interested in it. For example a person who wants to stay in a hotel may be interested to search for a feasible and good hotel to stay. This is possible either by checking each hotel one by one by him/her self or to listen about a suitable hotel from others. But finding people for verbal discussion is difficult. Therefore a fast and easy way is to get the point of view from others through internet. Internet is the fastest way to get user's opinion through blogs and reviews posted by others. Online products reviews are increasingly available and are frequently used by consumers to get the most suitable choice [2]. Manufacturers and market competitors are also getting advantage from these reviews to get, reputation feedbacks and compete in the market. According to a survey conducted by comScore with the Kelsey Group, reviews had a significant influence on purchase. They reported that 81% of Internet users (or 60% of Americans) have done online research on a product at least once and consumers that were willing to pay at least 20 percent more for services receiving an "Excellent," or 5-star, rating than for the same service receiving a "Good," or 4-star, rating [3]. For a popular product, the number of reviews can be in hundreds or even in thousands, which is difficult to be read them one by one. Therefore, automatic extraction and summarization of opinion is required [4].

The blogs and reviews are user generated text. This text is unstructured and unmanaged which needs proper arrangement to extract knowledge from it. Furthermore, not the whole text represents opinion but only a portion of a review or some sentences included opinion-oriented words. Thus opinion mining system needs only the required sentences to be processed to get knowledge efficiently and effectively. Automatic Opinion Mining (OM) is beneficial for both decision makers and ordinary people [5]. Automatic detection of opinionated and sentimental expression in text is becoming increasingly important from application point of view [6].

When we express some statement about something then we name some person or thing; and say some thing about that person or thing. In other words we must have subject to speak about and we must say or predicate something about that subject. Hence every sentence has two parts, the subject part which names the person or thing and the predicate part which tells something about the subject. e.g "the location was good" etc. Subjectivity is used to express private states in the context of a text or conversation. Private state is a general term for opinions, evaluation, beliefs, perception, emotions, peculation and etc [7]. If a user feedback has no judgment or opinion on the source content then it is called objective. Ahmed Abbas et al. [8] have presented a very good taxonomy about OM linguistic aspects. They have categorized the OM linguistic job as classification, features, techniques and domains. Actually when a user expresses opinion about a product then he/she states about the product as a whole or about its features one by one. The most important thing in classifying reviews documents is the choice of feature set [9]. Feature identification in product is the first step of opinion question answering and other opinion mining application [5]. Feature selection can be used for improving the efficiency and removing non-discriminating features [10]. Sometimes users express their opinions without explicit feature words. But we can still deduce the features on which their opinions focused from the review text. These kinds of features are Implicit Features for example, in the sentence "Grand hotel is fantastic" we can deduce that the feature which users talk about is grand hotel is about its facilities although the word

K. Khan is with Department of CIS, University Teknologi PETRONAS, Seri Askandar, PERAK, Malaysia (e-mail: khairullah_k@yahoo.com).

does not appear explicitly. The identification of implicit features is a harder task than identifying explicit ones [5].

We have tried to reach the subjective term through AV's. Our experiments prove that both subjective and predicates are linked through AV's. Thus AVs are good indicators for subjective terms in sentences. AVs are small number of seed words and can be applied to get the features and opinion-oriented terms efficiently and effectively. Through in-depth experiments we got the point that customer mostly discuss features and express their opinion about those features and more than 80% of sentences have any of the proposed AVs.. By combing our technique with dictionary-based extraction a generalized framework can be developed for opinion and sentiment extraction from documents. The rest of the paper is organized as follows. In section II we have discussed related work, section III presents methodology, and section IV represents experiments and results, while section V concludes the paper. For opinion extraction it is required to know the linguistic terms and get the idea from the text.

II. Related Work

In this section we review related work on the extraction of features of products and opinion-oriented words from text documents. Due to its interestingness and potential in the market, researcher has taken a keen interest in mining customer reviews. Different authors have worked on different aspects of this area. Some has worked on extraction of features and opinion-oriented word [11,12,13]. In [5,6] the authors have addressed the problem of scoring product reviews based on features of products from their textual documents. Minqing Hu and Bing Liu [4] have used features of the products for extraction of customer opinion. According to Livia Polanyi and Annie Zaenen, "The most salient clues about attitude are provided by the lexical choice of the writer, but the organization of the text also contributes information relevant to assessing attitude" [15]. Another main focus is on subjectivity detection. Changli Zhang et al. [16] in their work have used bag-of-word(BOW) and appraisal phrase and get 79.0% result through BOW and 80.26 with the combination of BOW and appraisal phrase. In [17] Xiaowen Ding and Bing Liu, by experiments have shown that context rules are helpful to improve the recall without much loss in precision. In [4] Minqing Hu and Bing Liu have used NLProcessor linguistic parser to parse each review to split text into sentences and to produce part of speech tags for each word like noun, verb, adjective etc. Some authors have taken term senses into account and assume that a single term can be used in different sense and can present different opinion. They use WordNet Synsets for different senses of the same term [18].

Most of the existing work has used linguistic parser and part-of-speech (POS) tagging to identify features and opinion-oriented words which is a lengthy process. We have proposed a novel method to identify features and opinion-oriented sentences from text documents. We classify the documents into two categories of sentences. The first category of sentences has AVs and while second category does not have AVs. This classification is simple and efficient because a limited seed words are to be scanned from the documents. Our experiments on huge dataset collected from TripAdvisor (one of the most popular online review sites for hotels and tourism activities) we found that the 82% of feature related sentences and 85% of opinion-oriented sentences have AVs. Furthermore, some sentences which do not have AVs can be converted in the structure of auxiliary sentences. e.g., "comfortable beds" can be written as "beds were comfortable", similarly the sentence "polite staff" can be written a "staff was polite" etc. Thus through a proper mechanism our proposed method can further be improved to reduce the processing time of classification. Another important aspect of our research is the extraction of both implicit and explicit features from review text. The identification of implicit features is a harder task than identifying explicit ones. Existing research on feature identification mainly focus on finding features that appear explicitly and domain-specific. In our proposed approach the AV's are general and can by applied to any domain and any type of features.

III. Methodology

A. Data Collection

We performed our experiments on the dataset of about 26000 hotel reviews downloaded from the site1 which is collected from TripAdvisor2 that is one of the popular review sites about hotels and traveling. The data is in XML format which contains different tags as shown in Fig. 1. We extracted only text of reviews using our own software module. After extraction we preprocessed the text to remove unnecessary words.

```
<review>
<id>16026844</id>
<title>"Relief at finding a good hotel in Rome"</title>
<text> this small hotel is in a fabulous location but the
    double room with two beds, shared with my sister, was
    tiny. there was barely space to move. the bedroom was
    huge compared to the mimiscule bathroom. there was a
    tiny shower stall, i mean so small a large man couldn't
    turn around and i couldn't raise my legs to shave them.
    breakfast was a bit of a disappointment. no fresh juice
    or fruit, just canned. nice bread but basic. i would not
    choose this hotel.</text>
<score>5/5</score>
<pros />
<cons />
<features />
</review>
```

Fig. 1 Sample review text

[1] http://patty.isit.cnr.it/~baccianella/reviewdata/
[2] http://www.tripadvisor.com/

B. Preparation of Training Set

We used 50 reviews to collect training sets of features and opinion-oriented words. Our training process was semi automatic. After removal of unnecessary word we calculated frequencies of each word. To shorten our process we selected only those words which have frequency greater than or equal to 2. Then these words were reviewed by two English language experts to remove unnecessary words and to separate features from opinion-oriented words. Thus we got a set of features and opinion-oriented words as shown in Table I and II respectively.

TABLE I
TRAINING SET FEATURES

Features	Features	Features	Features
apartment	front	restaurant	tv
area	garden	review	way
bar	guest	rome	window
bathroom	hotel	roof	apartments
bed	internet	rooftop	areas
bedroom	location	room	bars
beds	night	rooms	bathrooms
boutique	nights	shower	bedrooms
breakfast	people	site	beds
choice	piazza	staff	boutiques
city	place	stay	breakfasts
desk	price	steps	prices
distance	street	choices	reception
doors	showers	windows	tourists
fountains	sites	streets	tvs
gardens	door	suites	ways
guests	fountain	distances	places
locations	suite	restaurants	desks
piazza	tourist	reviews	

TABLE II
TRAINING SET OPINION-ORIENTED WORDS

OP Words	OP Words	OP Words	OP Words
Small	large	quiet	like
Great	fine	perfect	variety
clean	most	walking	continental
walk	near	excellent	decorated
good	pretty	short	noisy
nice	away	well	plenty
friendly	nearby	adequate	quite
helpful	warm	fantastic	really
comfortable	few	tiny	

C. Sentences Categorization

We categorized the sentences of each review into two groups by using simple rule-based approach. Group one includes those sentences which have any of the given AVs. We represented this category as sentences with auxiliary verbs (SAV). While in the other group all those sentences were included which were not in group one and called sentences without auxiliary verbs (SWAV). We processed 2143 reviews with 22203 sentences containing 75.25% SAV and 24.74% SWAV sentences as shown in Table III. Then applied training set and found total 9521 sentences having opinion-oriented words, within which 84.634% belong to category SAV while only 15.366 sentences belong to SWAV. Similarly we compared these sentences according to features-orientated words. By applying features training sets we found total 16712 feature-oriented sentences out of which 80.1999% belong to SAV and 19.8001% belong to SWAV.

IV. EXPERIMENTS AND RESULTS

The Table III presents the percentages of features and opinion-oriented words of each category. We found total 25793 feature-oriented words out of which 82.49% belong to SAV while 17.5045 belong to SWAV. We also calculated opinion-oriented words and found 13818 words out of which 85.93% belong to SAV while 14.0686 belong to SWAV. Thus we got 5 to 1 ratio in case of features while 6 to 1 in case opinion-oriented words which is in fact a big difference.

TABLE III
CLASSIFICATION ON THE BASIS OF OPINION-ORIENTED WORDS

	Features		Opinion-Oriented	
Category	Counts	Percentage	Counts	Percentage
SAV	22103	82.49	11874	85.93
SWAV	4690	17.50	1944	14.06
Total	26793		13818	

The Table IV shows the percentages of sentences of each category with opinion-oriented sentences. Over all we processed 22203 sentences out of which 75.25 contain AVs and only 24% sentences do not have AVs. The percentage ratio of opinion-oriented sentences to SAV is 84.634% while to SWAV is 15.366%. This is really a big difference and show that most of the opinion-oriented sentences include Avs.

TABLE IV
CLASSIFICATION ON THE BASIS OF OPINION-ORIENTED WORDS

	Over All Sentences		Opinion-Oriented Sentences	
	Count	Percentage	Count	Percentage
SAV	16708	75.25	8058	84.63
SWAV	5495	24.74	1463	15.36
Total	22203		9521	

In Table V we have calculated the percentages of sentences of each category with feature-oriented. Out of total sentences 16712 sentences have features. Out of which 80% of category SAV while only 19.800% of category SWAV sentences. This is clear from the table that SAV has high ratio of feature-oriented sentences.

TABLE V
CLASSIFICATION ON THE BASIS OF FEATURES

	Over All Sentences		Feature-Oriented Sentences	
Category	count	%age	count	%age
SAV	16708	75.25	13403	80.19
SWAV	5495	24.74	3309	19.80
Total	22203		16712	

The graph shown in Fig. 2 presents counts of review wise features in each category. The SAV sentences have higher counts of features in each review as compared to SWAV. From these results it is clear that the selected auxiliary verbs are good indicators for features and opinionated sentences.

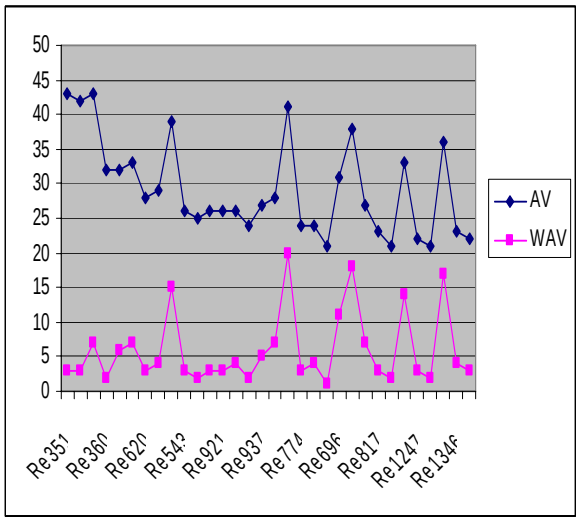

Fig. 2 Reviews against categories

We further investigated that which AV is mostly used in review. Table VI represents comparative analysis of AV used in each review. From this table it is clear that "has" is rarely used but sentence with it is most proven to feature oriented, while "are and is" are mostly used in opinion-oriented sentences.

TABLE VI
CLASSIFICATION ON THE BASIS OF FEATURES

	Sentences		Features		Opinion-oriented	
AVs	Counts	%age	Counts	%age	Counts	%age
are	1905	8.63	1618	84.93	1153	60.52
had	2191	9.93	1818	82.97	933	42.58
has	390	1.76	339	86.92	199	51.02
have	1652	7.49	1296	78.4	687	41.58
is	4868	22.07	4145	85.14	2929	60.16
was	7809	35.40	6463	82.76	4441	56.87
were	3241	14.69	2701	83.33	1719	53.03
Total	22056		18380		12061	

V. CONCLUSION AND FUTURE WORK

Features and opinion-oriented words extraction is the first step of opinion mining. In product reviews, users discuss features of products and state their views according to their experience and observations. Review text is unstructured and unmanaged which needs proper handling to extract knowledge from it. Furthermore not the whole text represents opinion but only a portion or some sentences include opinion-oriented words. Therefore opinion mining system needs only the required sentences to be processed to get knowledge efficiently and effectively. Actually when a user expresses an opinion about a product then he/she states about the product as a whole or about its features one by one. The most important thing in classifying reviews documents is the choice of feature set. We performed experiments and achieved good results by extracting features and opinion-oriented words from review text with help of auxiliary verbs. In future we will try to combine our idea with existing technique for extraction of features and opinion-oriented words

REFERENCES

[1] Bo Pang and Lillian Lee, "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval Vol. 2, Nos. 1–2 (2008) 1–135 2008, 2008.
[2] QI Su et al., Mining Feature-based Opinion Expressions by Mutual information Approach, International Journal of Computer Processing of Oriental Languages, 2007
[3] comScore, "Online consumer-generated reviews have significant impact on online purchase behavior," Press Release, http://www.comscore.com/press/release.asp?press=1928, November 2007.
[4]] Hu and Bing Liu, "Mining Opinion Features in Customer Reviews", American Association for Artificial Intelligence 2004, 2004.
[5] Changli Zhang1 et al., "Sentiment Classification for Chinese Reviews Using Machine Learning Methods Based on String Kernel", Third 2008 International Conference on Convergence and Hybrid Information Technology, 2008.
[6] Carlo Strapparava, Rada Mihalcea, "Learning to Identify Emotions in Text", SAC'08 March 1620, 2008, 2008.
[7] Nitin Jindal and Bing Liu, "Mining Comparative Sentences and Relations", American Association for Artificial Intelligence, www.aaai.org., 2006.
[8] Ahmed Abbas, etl. "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums" ACM Transactions on Information Systems, Vol. 26, No. 3, Article 12, 2008.
[9] Gamgarn Somprasertsri et al., Automatic product feature extraction from online product reviews using maximum entropy with lexical and syntactic features, in Proc. IRI 2008, 2008.
[10] Stefano Baccianella, et al., "Multi-facet rating of product reviews", ECIR 2009, LNCS 5478, pp. 461–472, 2009.
[11] A.M. popescu and O. Etzioni, extracting product features and opinion from reviews, in Proc. HLT-EMNLP 2005, 2005.
[12] M.Q. Hu and B. Liu Mining and summarizing customer reviews, in Proc. KDD(2004), 2004.
[13] N. Kobayashi, K. et al., Collecting evaluative expressions for opinion extraction, in Proc. IJCNLP(2004), 2004.
[14] V. Hatzivassiloglou and K. McKeown, "Predicting the semantic orientation of adjectives," in Proceedings of the Joint ACL/EACL Conference, pp. 174–181, 1997.
[15] Livia Polanyi and Annie Zaenen, "Contextual valence shifter", Computing Attitude and Affect in Text: Theory and Applications chapter 1, pages 1–10. Springer, 2006.
[16] Changli Zhang et al., "Sentiment Classification for Chinese Reviews Using Machine Learning Methods based on String Kernel", International on Convergence and Hybrid Information Technology, 2008.
[17] Xiaowen Ding and Bing Liu, "The Utility of Linguistic Rules in Opinion Mining", SIGIR'07, Amsterdam, Netherlands, 2007.

[18] Alina Andreevskaia and Sabine Bergler, “Mining Word-Net for fuzzy sentiment: Sentiment tag extraction from WordNet glosses”, In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL’06), pages 209–216, Trento, IT. 2006