International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Detecting Salient features and Summarizing Health Review using Latent Dirichlet Analysis

Mozibur Raheman Khan*, Rajkumar Kannan
Department of Computer Science, Bishop Heber College (Autonomous), Tiruchirappalli, India

ABSTRACT

Review means that "To examine something carefully especially before making decision or judgments". Health consumers especially for health service providers author health reviews. Since the number of reviews are enormous, hence there is need to summarize these reviews. In this paper, we propose a simple approach to select the interesting topics of health consumers discuss when reviewing their health providers online. Our approach does not rely on any manual tagging of the information, and operates on the text of online reviews. We analyze a large set of reviews and find out the topics discussed when reviewing providers with different specialties. The health-rating information is based on the sentiment-classification result. The condensed descriptions of health reviews are generated from the feature-based summarization. We propose a novel approach based on Latent Dirichlet Analysis (LDA) to identify health features. Furthermore, we find a way to reduce the size of summary based on the health features obtained from LDA.

Keywords: Health Consumers, Health Features, Latent Dirichlet Analysis, Natural Language Processing (NLP), Text Analysis, Text Mining

I. INTRODUCTION

In the health communication community, there is a widespread assumption that recent growth and advancement in net technologies such as (Web 2.0), notably the participative net (known as social media), have reworked the pattern of communication, as well as health-related communications. quite an sizable amount of people place confidence in the web as a supply of data and call aid for his or her health desires [1], there is a growing demand for automated tools, which might support the requirements of health shoppers on-line. One in every of the revolutions brought out by internet a pair of technology, is that the ability for internet users to place confidence in every other's opinions once creating selections starting from selecting a restaurants, buying or renting a house, dealing a film, to purchasing a laptop computer and so on. Currently each day the trust in health sector is step by step declining, hence there is a need to have associate degree opinion from the reliable sources. That is why the similar thought is applied to health domain, with a growing variety of internet sites dedicated to reviews of health practitioners authored by health shoppers. During a patient-centric apply, physicians have interest in understanding what matters to their patients once selecting a health supplier. Patients on their side would benefit from understanding what aspects of a practice of which other patients pay attention to when choosing a provider. For health researchers, it is essential to analyse what factors health consumers care about when assessing a provider, as it can influence health communication strategies. Finally, from a consumer health informatics standpoint, providing tools to process and organize the information conveyed in provider reviews can augment the functionalities of Personal

Health Records, provided the tools are accurate enough.

Earlier the label of satisfactions is well understood by surveys and questionnaires of patients, either to assess the effect of a particular element of the patient-provider interaction (e.g., [2]) or as a comprehensive analysis tool [3, 4]. Nevertheless, these factors are often established in a top-down fashion, from experts. Furthermore, because they are discussed in reports and papers for the scientific community, health consumers do not always rely on them when choosing providers. We propose the use of computational methods to conduct a complementary type of analysis: discovering key aspect that contributes to patient satisfaction and finding the corresponding opinion for the same aspect. By relying on the collective experience of health consumers as conveyed in the text of the reviews, we propose to identify the salient aspects about a provider that matter to the health consumers themselves.

While there has been much debate over the quality and impact of such source of information recently [5, 6, 7] (in particular the fear of fraudulent reviews and lack of trust in the authors), and much care has to be put into interpreting and using the results of any fully automated method of analysis, the phenomenon of peer reviewing seems to be a growing trend and a medium health consumers rely on more and more (if only measured as the ever-increasing number of reviews written by health consumers online and websites dedicated to this type of content) [8].This specific work depends on the reviews to spot trends through the text mining of huge amounts of reviews, thereby minimizing the impact of deceitful reviews.

Few websites provides a structured questionnaire for health consumers to review a health provider. For instance, the website HealthGrades1 allows for nine dimensions to be assessed, each on a 5-level scale: general recommendation (would you recommend this physician to friends and family), level of trust (do you trust the physician to make recommendations that are in your best interest), to which extent the physician helps patients to understand their condition, to which extent the physician listens and answers questions, the time spent with patient, the ease of scheduling, the office environment (cleanliness, etc.), the friendliness of the office staff, and finally the wait time. Other websites provide a hybrid of structured questions and free-text for reviewers to enter. The websites RateMDs2 and Zoc-Doc3, for instance, provide ratable dimensions (Zoc-Doc lists three dimensions: overall recommendation, bedside manner and wait time, while RateMDs lists four: punctuality, helpfulness, knowledge and overall recommendation) but also allow users to enter their own review. The variation over websites indicates that provider reviews is still an emerging genre of texts, with no set of standards for health consumers to follow. The fact that the genre is still fluid is advantageous for a quantitative, bottom-up analysis, as our goal is to discover salient points of discussion in reviews, without being influenced by a particular website's organization of information.

Figure 1. Screen shot with its reviews obtained from www.ratemds.com

Researchers in computational linguistics and information retrieval have investigated that how to identify aspects and sentiment from text automatically (see [9] for a complete review of techniques). However, most work to put up date has focused on product reviews (e.g., laptops, restaurants, movies). Recently generated reviews from health

consumer need to be processed for finding a hidden pattern to understand different aspects of health service provider. Applying computational methods to the analysis of reviews of health providers is timely and novel.

The major contributions of this paper are listed below:

- ✓ To identify the salient topics or aspects of health provider.
- ✓ Propose a novel approach based on LDA to identify health features. Health features and opinion words are used to select appropriate sentences to become a review summarization. Summarize the health reviews authored by health consumers.
- ✓ Propose an LDA-based filtering mechanism to allow the users to choose the features in which they are interested, and this mechanism could reduce the size of summary efficiently.

The rest of this paper is organized as follows. In Section II, related surveys are presented. In Section III, feature based review summarization is proposed. In Section IV, experiment and results is presented. In Section V, the conclusion is presented.

II. RELATED WORKS

With the growth and development of blogs, reviews and social networks opinion mining, review summarization and sentiment analysis became field of interest for many researches. Review summarization is Different from traditional text summarization and aims at producing a sentiment summary, which consists of sentences from a document that capture the author's opinion. The summary may be either a single paragraph or a structured sentence list. The former summary is generated by selecting some sentences or a whole paragraph in which the author expresses his or her opinion(s). The second one is generated by the auto-mined features that the author comments on. Our work is related to latter one.

Over the last few years, this special task of summarizing opinions has stirred tremendous interest amongst the Natural Language Processing (NLP) and Text Mining communities. 'Opinions' mainly include opinionated text data such as blog/review articles, and associated numerical data like aspect rating is also included. While different groups have different notions of what an opinion summary should be, we consider any study that attempts to generate a concise and digestible summary of a large number of opinions as the study of Opinion Summarization and is reported in [10].

The simplest style of an opinion outline is that the results of sentiment prediction (by aggregating the sentiment scores).The task of sentiment prediction or classification itself has been studied for several years. Beyond such summaries, the newer generation of opinion summaries includes structured summaries that provide a well-organized breakdown by aspects/topics, various formats of textual summaries and temporal visualization. The different formats of summaries complement one another by providing a different level of understanding. For example, sentiment prediction on reviews of a product can give a very general notion of what the users feel about the product. If the user need of more specifics, then the topic-based summaries or textual summaries would be more useful. Regardless of the summary formats, the goal of opinion summarization is to help users digest the vast availability of opinions in an easy manner. The approaches utilized to address this summarization task vary greatly and touch different areas of research including text clustering, sentiment prediction, text mining, NLP analysis, and so on. Some of these approaches rely on simple heuristics, while others use robust statistical models.

A. Aspect-Based Summarization

In general, aspect-based summarization is made up of three distinct steps - aspect/feature identification, sentiment prediction, and summary generation. Some approaches, however, integrate some of the

three steps into a single model. The feature identification step is used to find important topics in the text to be summarized. The sentiment prediction step is used to determine the sentiment orientation (positive or negative) on the aspects found in the first step. Finally, the summary generation step is used to present processed results from the previous two steps more effectively.

B. Aspect/Feature Identification

Various methods and techniques have been proposed to solve challenges in each of these steps. In the following three subsections, we will describe core techniques used in the aspect/feature identification step, the sentiment prediction step, and the summary generation step.

The process of mining opinions from Chinese review of products sold online described in [11].The structure of Chinese review is free, which leads to a more complicated relationship between opinions and features. Their papers introduce two main steps of opinion mining: feature extraction and opinion direction identification. Opinion mining and sentiment analysis actually focus on polarity detection and feature based opinion mining. These two discipline use data mining and natural language processing (NLP) techniques to discover retrieve and distill information and opinions from vast textual information. Many researchers attempt different techniques to detect the polarity of reviews. They extract "hot" features that a lot of people have comment in their reviews, and then finds those infrequent ones. In order to improve the accuracy of the experiment, redundant features are removed. The opinion direction identification function takes the generated features and summarizes the opinions into two categories: positive and negative. They consider adjectives and negative adverbs as opinion words and use the Naïve Bayes classifier to identify their direction. By opinion orientation, we mean whether an opinion is positive or negative.

In the recent work, shallow parsing was used to identify aspects for short comments in [12]. In short, comments, most opinions are expressed in concise phrases, such as 'well packaged' and 'excellent seller'. With this in mind, it is assumed that each phrase is parsed into a pair of head term and modifier, where the head term is about an aspect or feature, and the modifier expresses some opinion towards this aspect (e.g. 'fast[modifier]shipping[head]'). The head terms in the text are then clustered to identify k most interesting aspects.

There are different approaches introduced in [13].Their methods use a combination of text mining and econometric techniques.The methods attempt to first decompose product reviews into segments that evaluate the individual characteristics of a product (e.g., image quality and battery life for a digital camera).There is a slightly different approach for extracting features in movie reviews [14]. Since many of the features in their case are around the cast of a movie, they build a feature list by combining the full cast of each movie to be reviewed. A set of regular expressions is then used to identify whether a word in a review matched one of the words in the feature list in. A simple approach introduced to discover features [15]. They consider paragraph level frequencies as well as document level ones to help identify features.

Mining Techniques for Feature Discovery: Another commonly used method to identify features is a 'mining' approach [13, 16, 17 and 18]. Frequent item set mining can compensate the weaknesses of pure NLP-based techniques. This approach does not restrict that only certain types of words or phrases can become candidate features. Instead, other information like the support information is used to determine a particular word or phrase has feature or not. Certain non-promising features are even pruned with the use of mutual information and redundancy rules. This approach shows reasonable performance especially with product reviews.

Supervised association rule mining-based approach is used to perform the task of feature extraction in [17]. Their methods are based on the idea that each sentence segment contains at most one independent feature. First, each review sentence is divided into a set of sentence segments based on separation by '.', ',' , 'and', 'but', etc; then all the feature words are manually tagged. With the segmented and tagged data set, Association Rule Mining is performed to learn rules of the form A1 A2...An) [feature] for predicting feature words, based on the remaining words in a sentence segment and their POS tags. Since association rule mining does not account for the order of A1,A2...An in a sentence, many of the learnt rules can be pruned based on inconsistency of the patterns with English grammar. Features on a new input dataset are then extracted using these trained rules. In case two rules resulted in two different features for the same sentence segment, the more frequently occurring feature is chosen.

C. Sentiment Prediction

The feature discovery step is commonly followed by sentiment prediction on the text containing options that are found previously. Sentiment prediction in itself is an active research area. While there are many techniques solely for this task. In this section, we will discuss the techniques used within the framework of opinion summarization.

The standard machine learning outperforms human-proposed baselines is found in [19]. They employed naive Bayes, maximum-entropy classification, and support vector machines (SVMs) to perform sentiment-classification task on movie review data. According to their experiment, SVMs tended to do the best, and unigram with presence information turns out to be the most effective feature.

Some of the researchers in the recent years have extended sentiment analysis to the ranking problem, where the aim is to assess review polarity on a multipoint scale [20, 21, and 22]. The problem of analysing multiple connected opinions in a very text and conferred an algorithmic rule that put together learns ranking models for individual aspects by modelling the dependencies between assigned ranks self-addressed[22]. Graph-based semi-supervised learning algorithm to address the sentiment-analysis task of rating inference and their experiments showed that considering unlabelled reviews in the learning process can improve rating inference performance is proposed in[20].

One of the recent studies, using a learning-based strategy in aspect-based summarization in [12]. They propose two methods for classifying each phrase clustered into the k interesting aspects into a rating r(f). First they assume that the rating of each aspect is consistent with its overall ratings. In other words, each phrase mentioned in a comment shares the same rating as the overall rating of comments. With this assumption, the side ratings will be calculated by aggregating ratings of all the phrases concerning every side.

In the second method, instead of blindly assigning the same rate to each phrase as the overall rating of the comment, they learn aspect level rating classifiers using the global information of the overall ratings of all comments. Then each phrase is classified by the globally trained rating classifier. They essentially classify each phrase by choosing the rating class that has the highest probability of generating the modifier in the phrase, which is basically a Naive Bayes classifier with uniform prior on each rating class. The ratings are then aggregated by averaging the rating of each phrase within an aspect. This method of prediction is shown to work much better than just using the overall ratings.

Lexicon/Rule-based Methods for Sentiment Prediction: Lexicon-based sentiment prediction is very popular in the context of opinion summarization [14 ,15,17 and18]. This technique generally relies on a sentiment word dictionary. The lexicon typically contains a list of positive and negative words that are used to match words in the

opinion text. For example, if an opinion sentence has many words from the positive dictionary, we can classify it as having a positive orientation. These word lists are often used in conjunction with a set of rules or can be combined with the results of POS tagging or parsing.

For identifying the opinions about features and their orientation, [17,18] proposed a simple yet effective method based on WordNet. They start with a set of about 30 seed adjectives for each predefined orientation (positive and negative). Then they use the similarity and antonym relations defined in WordNet for assigning positive or negative orientations to a large set of adjectives. Thus, the orientation of an opinion of feature was decided by the orientation of the adjectives around it.

Similarly, the author has used a set of positive and negative words to predict sentiments [15]. They used two sets of sentiment words GI 2 and CNSD 3. They enlarged the seed vocabulary using two thesauri Cilin [J. et al. 1982] and BOW 4. The orientation of an opinionated sentence is decided based on the orientations of its words. Instead of employing a set of rules, they allotted sentiment scores to sentences allotted to topics. These scores represent the sentiment degree and polarity. Additionally to own a polarity of positive and negative, if bound words like 'say', 'present', and 'show' were gift within the sentence, a zero opinion score was allotted as a neutral opinion.

A dependency relationship is used to identify opinions associated with feature words in [14]. In order to identify the orientation of the opinions, they used a strategy similar to that of [17, 18].They identified the top 100 positive and negative opinionated words from a labeled training set and then used WordNet synsets to assign orientations to other words. Furthermore, the orientation of a word was reversed if there was a negation relation such as 'not' or 'anti' involved. This line of work is popular because it is simple and lexicons can be good features for learning-based methods. Lexicon-based approaches are known to work well in domains like product reviews where people are explicit about their expressions (e.g. 'The battery life is bad'). However, in harder domains like movie reviews where people are often sarcastic, such a method yields in poorer performance because the context was often ignored. Also, the performance of this method depends on the quality of the dictionary used. For the best performance, different dictionaries have to be defined for different domains and aspects.

Other Methods for Sentiment Prediction: words used in the vicinity of the features found as a starting point in predicting the sentiment orientation is proposed in[16]. Basic intuition is that an opinion phrase associated with a product feature tends to occur in its vicinity. Instead of using simple word window to check the words in vicinity, they use syntactic dependencies computed by MINIPAR [Lin 1998]. Heads and their corresponding modifiers in dependency parsing results are considered as potential opinion phrases. They then use a well-known computer vision technique, relaxation labelling [23], to predict the polarity of extracted opinion phrases. Relaxation labelling uses an update equation to re-estimate the probability of a word label based on its previous probability estimate and the features of its neighbourhood. The initial probability is computed using a version of Turney's PMI-based approach [24]. This technique is found to generate opinions and its corresponding polarity with high precision and recall. However, this is tested only on user reviews in the products domain, so it may not be general enough to be used in any arbitrary domain. In addition, since the sentiment prediction step alone is multi-faceted and very involved, the approach can have scalability issues.

Since our review is composed of sentences and a sentence is composed of terms, it is reasonable to determine the semantic orientation of the text from terms. As a result, the sentiment analysis research started from the determination of the semantic

orientation of the terms. Textual conjunctions such as "fair and legitimate" or "simplistic but well-received" to separate similarly connoted and oppositely connoted words is employed in [39].To determine the orientation of subjective terms based on the quantitative analysis of the glosses of such terms, i.e., the textual definitions that are given in online dictionaries is proposed in [40]. The process is based on the assumption that terms with similar orientation end to have "similar" glosses (i.e., textual definitions).Thus, synonyms and antonyms could be used to define a relation of orientation.

D. Summary Generation

Using the results of feature discovery and sentiment prediction, it is then critical to generate and present the final opinion summaries in an effective and easy to understand format. This typically involves aggregating the results of the first two steps and generating a concise summary.

In the following subsections, various summary generation methods for opinion summarization are described. While each technique has its own focus, some techniques can be combined with others. For example, we may add a timeline to text selection methods.

1) Statistical Summary: While there are various formats of summaries, the most commonly adopted format is a summary showing statistics introduced by [14,17, 18]. Statistical outline directly uses the processed results from the previous steps - a listing of aspects and results of sentiment prediction. By showing the quantity of positive and negative opinions for every side, readers will simply perceive the overall sentiments of the service providers.

The summarization statistics is displayed in a graph format. With the graph representation, we can obtain people's overall opinions about the target more intuitively [27]. Well-known software developed known as, Opinion observer, which shows statistics of opinion orientation in each aspect and even enables users to compare opinion statistics of several products, which compares opinions on three cell phones from three different brands [28]. This format of summary has been widely adopted even in the commercial world.

2) Text Selection: While statistical summaries help users understand the overall idea of people's opinion, sometimes reading actual text is necessary to understand specifics. Due to the large volume of opinions on one topic, showing a complete list of sentences is not very useful. To solve this problem, many of the recent studies try to show smaller pieces of text as the summary. They use different granularities of summaries including word, phrase and sentences level granularities. [12,15,16,29 and 30]

III. FEATURE BASED REVIEW SUMMARISATION

Figure2 provides the architectural overview of our proposed health reviews summarization system. The inputs for the system are a doctor's name and the salient features from the corresponding reviews. The output is the summary of the reviews as the one shown in figure5.The system performs the summarization in three main steps (as discussed before), the first step is to identify health features using LDA that have been commented on by health consumers; the second is identifying opinion sentences in each review and deciding whether each opinion sentence is positive or negative, and finally Summarizing the results. These steps are performed in multiple sub-steps.

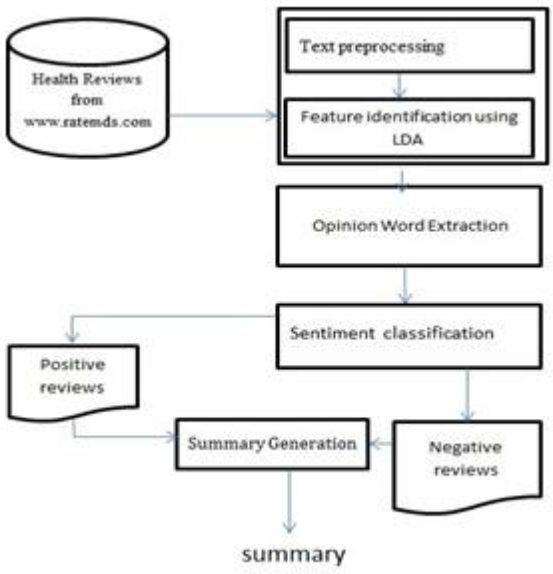

Figure 2. Architecture of Feature-based Reviews Summarization System

A. Data collection and Preprocessing

In order to create our datasets, we collected a corpus of reviews from the public RateMDs website. As a preprocessing step, the portions containing the reviews were extracted from the HTML pages, along with the specialty designation of each provider. The reviews were tokenized and separated to individual sentences. Stop words were removed. All the text documents combined is known as the corpus. To run any mathematical model on text corpus, it is a good practice to convert it into a matrix representation. LDA model looks for repeating term patterns in the entire Document Term matrix. Python provides many great libraries for text mining practices, "Gensim" is one such clean and beautiful library to handle text data. It is scalable, robust and efficient.

We stratified our dataset of reviews into six individual sets of reviews:
review of general practitioners(GP),obstetricians/ gynecologists(ObGyn),dentists(Dent),psychiatrists(Psych),dermatologist(Derm) and cardiologist(Card).We describe the main computational method on which we rely (Latent Dirichlet Analysis,or LDA) to identify salient aspects in reviews of health providers and how we customize it to answer our research questions. There are two challenges we address in particular: (i) dataset selection and selection of the unit of processing on which to apply LDA,and (ii) determining the optimal number of aspects discussed in the reviews (model order). We first give an overview of LDA in general,followed by our experimental setup.

B. Detecting salient feature using LDA

There are many approaches for obtaining topics from a text such as –Term Frequency and Inverse Document Frequency. NonNegative Matrix Factorization techniques. Latent Dirichlet Allocation is the most popular topic modelling technique proposed in this paper. LDA assumes documents are produced from a mixture of topics. Those topics then generate words based on their probability distribution. Given a dataset of documents, LDA backtracks and tries to figure out what topics would create those documents in the first place.

Our method for determining common topics discussed in medical reviews is based on a generative probabilistic graphical model, Latent Dirichlet Allocation (LDA) [31]. LDA is a fully unsupervised method to identify common topics of discussion in a collection of documents. The topics are identified automatically, without requiring any prior knowledge or manual annotation. This is particularly attractive to our task, since we want to discover the common topics discussed in reviews of health providers, rather than making hypotheses about the aspects of a health provider practice that are important to health consumers and validating them through data analysis.

We can get an idea of its subject, and a label can be assigned. The generative nature of the model allows it to handle newly observed documents which do not conform precisely to a previously seen distribution. Comparing LDA to other models proposed in the literature, and report improved results on document modelling and text classification tasks, where their model does considerably less over-fitting than the others [31]. Since then, LDA has been applied to many tasks, such as entity resolution [32],

information retrieval [33] and image processing [34]. Several efficient methods have been developed for inference with LDA. In this work, we employ a standard implementation of LDA which uses Gibbs sampling for parameter estimation and Inference.

1) LDA for Reviews of Health Providers: A specially tailored model [35], based on LDA, was shown to be effective at finding rateable aspects of hotel reviews, with the help of additional aspect-specific information provided by the reviewers. In [36], the authors demonstrated that a local version of LDA, which operates on individual sentences rather than documents, and doesn't require additional information, can find ratable aspects in a variety of domains including product and restaurant reviews. We hypothesize that a similar approach would be suitable for the domain of professional services and, in particular, for our task of determining the salient aspects in online reviews of health providers.

2) Model Order: The issue of model order, i.e., determining the correct number of clusters (in our case the discovered topics), is an important element in unsupervised learning. A common approach [37, 38] is to rely on a cluster validation procedure. In such a procedure, different model orders are compared, and the one with the most consistent clustering is chosen. For the purpose of the validation procedure, we have a cluster corresponding to each aspect, and we label each sentence as belonging to the cluster of the most probable aspect.

Given the collection of sentences in our data, D, and two connectivity matrices C and $\hat{C}$, where a cell i, j contains 1 if sentences d_i and d_j belong to the same cluster, we define a consistency function F (following [38]):

$$F(C, \hat{C}) = \frac{\sum_{i,j} 1\{C_{i,j} = \hat{C}_{i,j} = 1, d_i, d_j \in \hat{D}\}}{\sum_{i,j} 1\{C_{i,j} = 1, d_i, d_j \in \hat{D}\}} \qquad (1)$$

The algorithm for LDA is as follows

1. Run the LDA model with k topics on D to obtain connectivity matrixC_k.
2. Create a comparison connectivity matrix R_k based on uniformly drawn random assignments of the instances.
3. Sample random subset D^i of size $\delta|D|$ from D.
4. Run the LDA model on D^i to obtain connectivity matrixC^i_k.
5. Create a comparison matrix R^i_k based on uniformly drawn random assignments of the instances in D^i.
6. Calculate $score_i(k) = F(C^i_k, C_k) - F(R^i_k, R_k)$
1. Where F is given in Eq. 1.
7. Repeat steps 3 to 6 q times.
8. Return the average score over q iterations.

This procedure calculates the consistency of our clustering solution, using a similar sized random assignment for comparison. It does this on q subsets to reduce the effects of chance. The k with the highest score is chosen. In our experiments, we used q = 10, δ = 0.9, and let k range from four to fifteen.After a number of iterations, a steady state is achieved where the document topic and topic term distributions are fairly good. This is the convergence point of LDA.

C. Opinion-Words Extraction

In addition to feature identification, opinion words about the product features are important as well. The opinion words by retrieving the nearby adjective of product features are proposed in [39]. In addition to language sentence-structure characteristic, the use of dependency grammar graph to find out some relations between feature words and the corresponding opinion words in training data [40]. They both rely on language sentence structure to extract opinion words; therefore, these approaches will be applicable to those language sentences having such a characteristic. Many languages do not possess the aforementioned sentence structure. Hence, we propose to use a statistical approach to discover opinion words. First, we take into account POS-tagging information of the opinion words. According

to their analysis, adjectives are usually used to describe sentiment; therefore, these terms become the candidate opinion words. Second, term frequency is taken into account; therefore, frequency of the opinion words should exceed a threshold value. Let AVG be the average of sum of square of frequency of all items as shown in (3) below. A $term_i$ will be selected only if its square of frequency is equal or larger than AVG. We manually selected positive and negative sentences from 500 positive reviews and 500 negative reviews, respectively. Positive opinion words and negative opinion words could be further obtained based on term frequency and POS tagging.

$$S_f = \sum_{i=0}^{n} \{\text{Frequency}(term_i)\}^2$$

$$AVG = S_f / n. \qquad (3)$$

D. Feature-Based Summarization:

In general, feature-based summarization is based on medical related features and opinion words. It is not easy to use compression ratio directly, since the sentence-selection criterion is based on the presence of medical/health features. Hence, we propose an LSA-based filtering approach to further select the content of the summary based on user's favor. In health organization we are interested in finding health feature from the health reviews and we employ LDA to find out related feature terms of a specific health feature, and these related terms could be regarded as being semantically related to the health organization. For each given product feature f, LDA could discover related terms F that are semantically related to f. In general, F could be regarded as f's related terms, and the system can employ F to select summary sentences. In application design, the system provides all the summary sentences in the beginning. The product-feature seeds mentioned in LDA-based feature-identification process will become candidate interested summarization features. The system allows the user to determine the feature f in which he/she is interested. When the user determines f, the system will generate a summary, which is related to health domain.

Practically, a positive health review may include negative comments about specific aspects and vice versa. In this paper, we propose to analyse the polarity of a movie review using logistic regression and analyze the polarity of a sentence using opinion words. In feature-based summarization, the system can utilize the polarity of opinion words to determine the polarity of sentences. Hence, the system can provide both positive- and negative-review summarization, regardless of the polarity of a review.

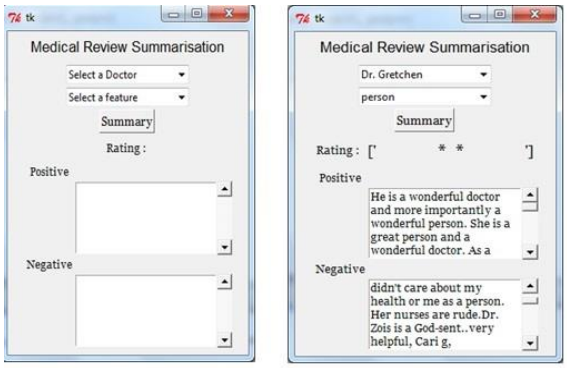

Figure 3(A) **Figure 3(B)**

Figure 3, (A) Summarization screen shot **(B)** Screen shots with review

IV. EXPERIMENTS AND RESULTS

We have performed several experiments to evaluate our system. In sentiment-classification experiment, logistic regression is employed to perform the sentiment-classification task. Several feature combinations are used to evaluate the system performance. To identify the health feature from medical, we propose an LDA based approach to identify the health features authored by health consumers.

A. Data Set:

We have collected health reviews of five hundred doctors from ratemds.com and these reviews have been placed in reviews database. This site provides

hundreds of reviews for thousands of doctor from across the globe. Each of the reviews includes a text review and other numeric ratings are available for various other features. We have received all these from family doctor/GP, Gynecologist, podiatrist and orthopedic and so on. The site provides numerical rating of four aspects namely staff, punctuality, helpfulness and knowledge. Textual comments are written by the health consumers with an average of three sentences. For each doctor, we first downloaded the first available reviews. Looking at the sites nearly we can understand that there are ten important specialty available. They are Internist, Gynecologist, /general, podiatrist, Dentist, Psychiatrist, Orthopedist, Cardiologist, Gastroenterologist, and Dermatologist and so on. For each specialty there are top reputed doctors are available and each doctor is receiving hundreds of reviews. For five hundred doctors, we have collected 5000 reviews and approximately 20000 features included in this paper.

B. Salient Feature Identification:

In the effort of discovering what health consumers consider salient aspects when reviewing providers, we had a set of desiderata for our computational methods: dynamic and bottom-up, without any reliance on manual annotation. Our results show that LDA is an appropriate method given our constraints. Furthermore, when reviews are processed at the sentence level (rather than as a whole), and reviews are grouped by specialty, it is possible to identify salient aspects that are specialty-specific. The discovered aspects which are common to all specialties resemble the traditional aspects of patient satisfaction questionnaires (such as bedside manner of the staff and the provider, and level of attention provided by the provider to the patients). When examining the aspects that are specific to different specialties, however, interesting patterns emerge. For instance, cost is a salient topic only for dentists. This makes sense, as while most reviewers have medical insurance, coverage for dental procedures is less common, and cost becomes a salient topic. Similarly scheduling is particularly salient for ObGyns and psychiatrists, but not for other specialties.

Table 1. Top ten Features identified using LDA

Care
care patient
Recommend
office time
like staff
Visit
friendly helpful
best ever
recommend highly
care patient
thorough caring

Table 2. Precision – Recall of Logistic Regression, SVM and Gaussian Naïve Bayes Classifiers. Experiments with reviews on doctors (using top 85 frequent features)

Classifier	Sentiment	Precision	recall	F1-score
Logistic Regression	-1	0.80	0.21	0.33
	0	0.68	0.93	0.78
	1	0.75	0.43	0.55
	Average	0.71	0.69	0.66
SVM	-1	0.60	0.32	0.41
	0	0.69	0.90	0.78
	1	0.65	0.40	0.50
	Average	0.67	0.68	0.65
Gaussian Naïve Bayes	-1	0.18	0.89	0.29
	0	0.80	0.14	0.24
	1	0.32	0.26	0.29
	Average	0.58	0.27	0.26

C. Sentiment Classification:

Opinions in natural language are usually expressed in subtle and complex ways. For example, the polarity of a sentence may be changed when a negative term

is used in the sentence. We considered possible feature combination in the experiments to obtain the best feature selection. Based on the bag-of-words model, we used unigram, bigram, negation, location, frequency, and presence features (i.e., only consider whether the feature is present or not) to perform the classification task with different feature combinations. In feature selection, our experiments also showed that unigram with presence features outperforms bigram with other features, and the result is the same as described in [19]. In addition to unigram with presence features, we design three basic experiments to compare the differences of feature combinations, and they are described as follows.

1) Type I: a) Removal of the terms appearing in both positive and negative reviews;
b) frequency-feature criterion, where the term's square of frequency should be at least AVG, as shown in (3);
2) Type II: frequency-feature criterion, where the term's square of frequency should be at least AVG, as shown in (3);
3) Type III: frequency-feature criterion, where the term should occur at least three times.

The Type I experiment includes two additional features to evaluate its performance. The first feature is about the removal of the terms appearing in both positive and negative reviews. In general, the terms that appear in both positive and negative reviews could not provide enough semantic orientation to differentiate positive and negative reviews. The second feature is about the comparison of the effect of frequency.

The Type I and Type II experiments are used to compare the effect of term selection. While Type I removed the terms appearing in both positive and negative reviews, the Type II experiment used all the terms. The Type I and Type III experiments are used to compare the effect of term frequency. While Group 2 used the frequency criterion based on (3), Type 1 selected the terms that occur at least three times.

These three experiments are performed to evaluate their performances on movie-review data, and they will become the bases of other experiments. Negation and position are additional features that are included into these three bases to perform feature combination. In negation feature, a negation term may change the polarity of a sentence completely, which may blur the decision. For an example, a sentence "This movie is interesting" indicates a positive opinion about this movie, while the sentence "This movie is not interesting" changes the polarity of the sentence. As for position feature, people may have the conclusion in the end; therefore, position feature is employed, as well to evaluate its effect.

Table 2 shows the experimental result. Unigram with presence feature (i.e., only considers the presence and absence of a term) outperforms the other feature combinations, and this result conforms to result [14]. It seems like that negation, location, and bigram features do not contribute to sentiment classification. If we compare the performance of three basic experiments, type II outperforms type I and type III. In other words, the removal of the terms appearing in both positive and negative reviews will decrease the classification-accuracy rate. Meanwhile, the frequency criterion based on (3) is a little better than the frequency criterion, which is at least three times. Furthermore, the feature-combination experiments show that type II with negation feature outperforms type II, and this result is different from [14] research result.

However, sentiment-classification accuracy is not the only issue on mobile platform, and response time should be considered as well. Table IV shows that the system using unigram with presence feature will have 40 000 features, and it takes about 120 s to load the classification model. Obviously, it is infeasible on mobile platform if a system's response takes 120 s. Hence, the number of features is crucial to the

system's response time. We employ frequency as filtering criterion to reduce the number of features. The number of features could be reduced to 100 if we use the frequency criterion based on (3). Table IV shows that it takes about 6 s to load classification model, and it is feasible on mobile platform. Therefore, this frequency criterion is employed to perform sentiment classification only.

Table 3. Experiment Result for Different Feature Combinations

Features	Accuracy
Unigram with presence of feature	85.40%
Type I	71.00%
Type I Group1+ negation	70.79%
Type II	78.46%
TypeII + negation	79,32%
TypeII + position	71.64%
TypeIII	76.55%
TypeIII + negation	75.48%
Type III + position	70.15%

Table 4. SVM Model Loading and Prediction Evaluation Result(Sec)

Feature type	Number of features	Model loading	prediction
Frequency-based	100	5.25	< .0625
Unigram with presence	40000	120	0.5-0.625

The performance of sentiment classification on another movie review dataset, which is available at http://www.cs.cornell.edu/People/pabo/movie-review-data/. The dataset includes 1000 positive and 1000 negative movie reviews. Similarly, SVM is used to perform the classification task. The kernel function used in the system is RBF and K-fold cross validation (i.e., K = 5) is used in the experiment. Different feature-selection criteria are used in the experiment to compare their number of features and accuracies. Table V shows the experimental result, which includes three feature-selection approaches. The pre-processing task includes the punctuation-elimination process, the lowercase-conversion process, and the negative-term conversion process, which converts "n't" to "not." The first one used all the unigrams as features, while the second one employed frequency as the filtering criterion, with only the unigrams with occurrences more than three would be taken into account. The third one employed the frequency criterion listed in (3). The term-document matrices of all the experiments employed unigram with presence feature as entry value. The first two approaches do not remove stop words, but the third one removes stop words first. The main reason is that stop words are the terms with high frequencies; therefore, almost only stop words will be left using the criterion listed in (3) if the stop words are not removed in advance of the process.

Table 5. Sentiment-classification results using public movie-review dataset

Feature selection criterion	No. of features	accuracy
Unigrams	30, 084	86.5%
Unigrams with occurance more then 3	15,026	86.25%
Unigrams using frequency criterion baesd on eq.(3)	861	81.2%

The experimental results are similar to the previous experiment. The first one outperforms the other ones, but the number of features is enormous. The second one can reduce more than half of the features and the accuracy is almost the same. However, the number of features is still enormous. The number of features in the third experiment is 861 and its accuracy is about 81.2%. Although the accuracy of the third one is not as good as the other ones, it can dramatically reduce the number of features. Meanwhile, its accuracy is still acceptable practically.

D. Summary Generation:

Dr. Gretchen
Feature Recommend
Rating *

Positive:
<We highly recommend him>
<Dr. Liddell is wonderful and I recommend him highly to my friends and family>
<She even remembers past conservations we've had! Appointments are readily available but am sure once word gets out how good she is, it will get harder! Highly recommend Dr. Bortolotti.>
<I finally found my Doctor! Took 20 years!!!!She never rushes you out of her office, and if you call to speak to her, SHE calls you back. (instead of a nurse) I would highly recommend!!I highly recommend her. I highly recommend her to everyone.>
….
Negative :

<The only complain is long wait to see her.>
<With that said I highly recommend her...She doesn't just go "by the book.>
<" I highly recommend her!I really like him and his staff, but have had some trouble with getting prescriptions filled in a timely manner, which I found frustrating, but was only an issue because I was in and out of town (and may have been>
...

Figure 5. Review Summarization of a Health Service Provider

E. Discussion:

The results of topic models area unit utterly passionate about the options (terms) gift within the corpus. The corpus is delineated as document term matrix, which generally is incredibly distributed in nature. Reducing the dimensionality of the matrix can improve the results of topic modelling. Supported my sensible expertise, there are few approaches which do the trick.

Sometimes LDA may be used as feature choice technique. Take an example of text classification problem where the training data contain category wise documents. . If LDA is running on sets of class wise documents, Followed by removing common topic terms across the results of various classes can offer the simplest options for a class.

This study has a few limitations. While the use of LDA has been validated in several settings as an accurate tool for identifying topics of discussion in a large corpus of documents [31, 32, 33]. In this study only a shallow manual review of the topics was mentioned. In our future work, we plan to conduct a more in depth validation of the topics with the help of public health experts. Another limitation concerns the dataset: in our experiments, we selected reviews from a single website. Our methods can scale to a larger number of reviews and reviews from different websites. As such, this is often a limitation of our experimental setup, instead of the tactic itself.

V. CONCLUSION

In this paper, we have a tendency to style and implement a medical review-summarization system and Sentiment classification is applied to the medical reviews. We present a method to identify the salient aspects discussed in reviews of health providers authored by health consumers online. While there has been abundant work on the event and the chemical analysis of questionnaires to assess the factors referring to patient satisfaction, this work takes a complimentary approach and LDA is proposed to identify the salient aspects that health consumers care about when choosing a health provider in a quantitative manner. The aspects are learned automatically from a collection of reviews entered by health consumers, without any information other than the text of the reviews. In feature-based summarization, feature identification plays an essential role, and we propose a novel approach based on LDA to identify related health

features. Moreover, we use a statistical approach to identify opinion words. Health features and opinion words will be used as the basis for feature-based summarization.

VI. REFERENCES

[1]. W. Chou, Y. Hunt, E Beckjord, R Moser, and B. Hesse." Social media use in the United States: Implications for health communication". J MedInternet Res, 11(4):e48, 2009.

[2]. L Frostholm, P Fink, E Oernboel, K Christensen, T. Toft, F Olesen, and J .Weinman. "The uncertain consultation and patient satisfaction: The impact of patients' illness perceptions and a randomized controlled trial on the training of physicians' communication skills". Psychosomatic Medicine,67:897–905, 2005.

[3]. H Rubin, B Gandek, W Rogers, M Kosinski, C McHorney, and J Ware. Patients' ratings of outpatient visits in different practice settings results from the medical outcomes study. JAMA,270(7):835–840, 1993.

[4]. Press Ganey Associates. Medical practice pulse report: Patient perspectives on American health care. http://www.pressganey.com/galleries/defaultfile/2009 Med Practice PulseReport.pdf, 2009.

[5]. S O'Brien and E Peterson. Identifying high quality hospitals: Consult the ratings or flip a coin? Arch Intern Med, 167(13):1342–1344,2007.

[6]. B Hughes, I Joshi, and J Wareham. Health 2.0 and Medicine 2.0: Tensions and controversies in the field. J Med Internet Res, 10(3):e23, 2008.

[7]. G Eysenbach. Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. J Med Internet Res, 10(3):e22, 2008.

[8]. G Eysenbach. From intermediation to disintermediation and apomediation: new models for consumers to access and assess the credibility of health information in the age of Web 2.0. Stud Health Technol Inform, 129(Pt 1):162–166,2007.

[9]. B Pang and L Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1–135, 2008.

[10]. Kim, Hyun Duk, et al. "Comprehensive review of opinion summarization." (2011).

[11]. Ge, Wang, Pu Pengbo, and Liang Yongquan. "Feature Extraction and Opinion Summarization in Chinese Reviews." Open Automation and Control Systems Journal 7 (2015): 533-539.

[12]. LU, Y., Zhai, C., and Sundaresan, N. 2009. Rated aspect summarization of short comments. In WWW '09: Proceedings of the 18th international conference on World wide web. ACM, New York, NY, USA, 131–140.

[13]. Archak, N., Ghose, A., and Ipeirotis, P. G. 2007. Show me the money!: deriving the pricing power of product features by mining consumer reviews. In KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York, NY, USA, 56–65.

[14]. Zhuang, L., Jing, F., and Zhu, X.-Y. 2006. Movie review mining and summarization. In CIKM '06:Proceedings of the 15th ACM international conference on Information and knowledge management. ACM, New York, NY, USA, 43–50.

[15]. KU, L.-W., Liang, Y.-T., and Chen, H.-H. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW).100–107.

[16]. Popescu, A.-M. and etzioni, O. 2005. Extracting product features and opinions from reviews. In HLT '05:Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Morristown, NJ, USA, 339–346.

[17]. Hu, M. and Liu, B. 2004a. Mining and summarizing customer reviews. In KDD '04:

Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York,NY, USA, 168–177.

[18]. Hu, M. and Liu, B. 2004b. Mining opinion features in customer reviews. In AAAI'04: Proceedings of the 19th national conference on Artifical intelligence. AAAI Press, 755–760

[19]. B. Pang, L. Lee, and S.Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in Proc. ACL-02 Conf. Empirical Methods Natural Lang. Process., 2002, pp. 79–86.

[20]. B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in Proc. 43rd Annu.Meet. Assoc. Comput. Linguist, Morristown, NJ: Assoc. Comput. Linguist.,2005, pp. 115–124.

[21]. A.B. Goldberg and X. Zhu, "seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization," inProc. TextGraphs: First Workshop Graph Based Methods Nat. Lang.Process, Morristown, NJ: Assoc. Comput. Linguist., 2006, pp. 45-52.

[22]. B. Snyder and R. Barzilay, "Multiple aspect ranking using the good griefalgorithm," in Proc. HLT-NAACL, 2007, pp. 300–307.

[23]. Hummel, R. A. and Zucker, S. W. 1987. On the foundations of relaxation labeling processes. 585–605. J., M., Y., Z., Y., G., AND H., Y. 1982. tong2yi4ci2ci2lin2. Shanghai Dictionary Press.

[24]. Turney, P. D. and Littman, M. L. 2003. Measuring praise and criticism: Inference of semantic orientation from association. ACM Trans. Inf. Syst. 21, 4, 315–346.

[25]. V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in Proc. 8th Conf. Eur. Chap. Assoc. Comput.Linguist., Morristown, NJ: Assoc. Comput. Linguist., 1997, pp. 174–181.

[26]. A. Esuli and F. Sebastiani, "Determining the semantic orientation of terms through gloss classification," in Proc. 14th ACM Int. Conf. Inf. Knowl.Manage., 2005, pp. 617–624.

[27]. Hu, M. and Liu, B. 2006. Opinion extraction and summarization on the web. In AAAI'06: proceedings of the21st national conference on Artificial intelligence. AAAI Press, 1621–1624.

[28]. Liu, B., Hu, M., and Cheng, J. 2005. Opinion observer: analyzing and comparing opinions on the web. In WWW '05: Proceedings of the 14th international conference on World Wide Web. ACM, New York, NY, USA,342–351.

[29]. Titov, I. and Mcdonald, R. 2008. Modeling online reviews with multi-grain topic models. In WWW '08:Proceeding of the 17th international conference on World Wide Web. ACM, New York, NY, USA, 111–120.

[30]. Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In WWW '07: Proceedings of the 16th international conference on World Wide Web.ACM, New York, NY, USA, 171–180.

[31]. D Blei, A Ng, and M Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research,3:993–1022.

[32]. I Bhattacharya and L Getoor. A latent dirichlet model for unsupervised entity resolution. Proc.SIAM International Conference on Data Mining.2006.

[33]. X Wei and W Croft. LDA-based document models for ad-hoc retrieval. Proc. of the ACM SIGIR conference. pp. 178-185. 2006.

[34]. L Fei-Fei and P Perona. A Bayesian hierarchical model for learning natural scene categories. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2005.

[35]. I.Titov and R McDonald. A joint model of text and aspect ratings for sentiment summarization.Proc. of the Conference of the Association for Computational Linguistics (ACL). pp 308–316.2008.

[36]. S Brody and N Elhadad. An unsupervised aspect-sentimentmodel for online reviews.

Proc.of the Conference of the North American Chapter of the Association for computational Linguistics(NAACL-HLT), pp 804–812, 2010.

[37]. E Levine and E Domany. Resampling method for unsupervised estimation of cluster validity. Neural Comput., 13(11):2573–2593, 2001.

[38]. ZY Niu, DH Ji, and CL Tan. I2R: Three systems for word sense discrimination, Chinese word sense disambiguation, and English word sense disambiguation. Proc. of the International Workshop on Semantic Evaluations (SemEval), pp.177–182, 2007.

[39]. M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proc.10th CMSIGKDD Int. Conf.Knowl. Discov.DataMining, 2004, pp. 168–177.

[40]. L. Zhuang, F. Jing, and X.-Y. Zhu, "Movie review mining and summarization,"in Proc. 15th ACM Int. Conf. Inf. Knowl. Manage., 2006, pp. 43–50.