© 2018 IJSRST | Volume 4 | Issue 5 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Summarizing Health Review using Latent Semantic Analysis

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

ABSTRACT

The amount of reviews is written by health consumer for health service supplier is growing every day. Text summarization reduces info as a shot to alter users to seek out and perceive relevant services of a health service supplier additional quickly and effortlessly. During this paper, we tend to propose a health review-summarization system based on features. The health-rating information relies on sentiment-classification results of the reviews. The feature-based health summarizations are generated from the reviews of health provider. We tend to propose a completely unique approach supported latent semantic analysis (LSA) to spot health options. What is more, we've got reduced the dimensions of outline supported the health options obtained from LSA. We have considered bottom-up approach for reviews collection and this approach provides a better reliability among health consumers. We think about each sentiment-classification accuracy and system latent period to style the system. The summarization of health reviews can be applied to the reviews of different service providers. Recent years have witnessed a significant growth to analyse the reviews and techniques have been developed to judge numerous summarization techniques in various domain. The goal of this paper is to provide short summaries of health reviews authored by health customers for varied health service suppliers.

Keywords: Health Rating, Latent Semantic Analysis, Bottom-up Approach, Health Consumers.

I. INTRODUCTION

In specific, on-line opinions is playing a unique role in generating virtual currency for businesses desirous to plug their merchandise, identifying the new prospects in business, and it helps us to manage the best standards. Cellular phones have completely become the most-vital a region of our lives. Mobile platform is presently one of the foremost customary platforms inside the globe used for exchanging any reviews notwithstanding the placement of the service supplier. However, the digital information displayed in cellular phones is impermissible in size, since cellular phones physically very little. Hence, an appropriate mechanism is framed that can provide users with condensed descriptions of documents will facilitate the delivery of digital content in cellular phones. This paper explores a system for health review report throughout that linguistics orientation of comments, the limitation of very little show capability of cellular devices, and the system time is considered.

Practically, after we are not conversant in a selected health service supplier, we have a tendency to raise our trusty sources to suggest one. Whenever we want to form a choice to go to any health service supplier, we have a tendency to sometimes raise different people's opinion. Since their call involves defrayment time or/and cash, what others suppose receives nice significance. Today, the recognition of the net drives folks to look for others opinions from the internet community before shopping for a product or seeing a movie. Web sites provide user rating and genuine comments on various services and these reviews could replicate users' opinions a few of product and different array of services. Consider a simple example, the health consumers authored reviews for various provider in ratemds.com and it lists the amount of reviews, various stages of ratings and, and comments from reviewers. If a patients needs to visit the hospital or a person desires to purchase a books, laptops or mobiles; these comments and opinions mostly influence their purchasing behaviours. Apart from these sites, a search engine is another important reliable source of information to supply for individuals to go booking for alternative people's opinions. The concerned question is typed by the user into a search engine; the computer program checks its index and provides an inventory of best-matching web content in keeping with its criteria, sometimes with a brief outline containing the document's title and, sometimes, components of the text.

Significant advances are achieved recently in text report. As a result, several applications that leverage text report techniques became out there to the general public. There has been a growing interest in research community in text summarization techniques in the biomedical domain[1].A Literature survey was conducted by Afantenos et al. identified ten most interesting biomedical text summarization studies appeared between 1999 and 2003.Since then; there are very important advances inside the report tools and techniques utilized within the medical specialty domain [2].

In recent years, the matter of "opinion mining" has seen increasing attention[3-5]. With the proliferation of reviews, ratings, recommendations, and various styles of on-line expression, on-line opinion may offer necessary information for businesses to market their merchandise, confirm new opportunities and understanding its prospectus, and moreover it helps to manage reputations of an individual. For example, most recommendation systems plan to alleviate information overloaded by distinguishing that things a user can realize worthy, and collaborative filtering is used rather than content filtering depends on the opinions of comparable customers to advocate items[6]. Primarily, the task of deciding whether or not or not a health review is positive or negative is corresponding to the quality binary-classification. If a review is given, then the classifier attempts to classify the review into positive class or negative class. However, opinions in language area unit sometimes expressed in refined and sophisticated ways. Thus, the challenges may not be self-addressed by simple text-categorization approaches like n-gram or keyword identification approaches[7].

In this paper, we have proposed feature based health review summarization using LSA technique applied to the medical reviews. The following is the main contributions of this paper:

- Propose a summarization health review system is suitable even in a mobile environment. We considered system response time issue to design the mobile application, and also the same system style will be extended to different domains with somewhat modification.
- Propose a unique approach supported LSA to spot salient options of health supplier. Health options and opinion words square measure won't choose applicable sentences to become a review report.
- To provide the genuineness of all health service providers, we have discovered the aspects automatically in a bottom-up fashion from the text of the health reviews authored by health consumers.

- Users are allowed to choose the features in which he/she is interested and this mechanism could reduce the length of summary drastically.

The rest of the paper is organized as: related work is presented in section II. Identifying the salient features using LSA techniques is presented in section III. Section IV represents

Feature based review summarization. Several experiments are introduced in section V. Conclusion is presented in section VI.

II. RELATED WORK

Shallow parsing was accustomed establish aspects for brief comments[8].In short comments; most of the opinions square measure expressed in summary phrases, like 'well packaged' and 'excellent seller'. With this in mind, it's assumed that every phrase is parsed into a try of head term and modifier, wherever the pinnacle term is concerning a facet or feature, and also the modifier expresses some opinion towards this side (e.g. 'fast[modifier] shipping[head]'). The pinnacle terms within the text square measure then clustered to spot k most interesting aspects.

There are different approaches introduces in[9]. Their ways use a mixture of text mining and economic science techniques. The initial plan to decompose health reviews into segments that value the individual characteristics of a health service provider (e.g., expertise of a doctor, approach of a doctor and other medical facility of a service provider).Then they adopt different ways but specifically the hedonistic [Rosen 1974] regression construct: (a) To find the burden that consumer place on every individual feature (b) to find the implicit analysis score that customers assign to every feature, and (c) however these evaluations have an effect on the revenue for a given product. By mistreatment product demand as an objective performs, they derive a context-aware interpretation of opinions. Supported the analysis, they show however customers interpret the announce comments and the way the comments have an effect on customers' selections. The intuition here is that the results are often utilized by the health service provider to see that which options contribute most to the demand for his or her services. Such info also can facilitate the service provider changes in quality of service over the course of a product's life cycle.

The approaches adopted by researchers may be classified in two main

categories, the first one is machine learning and the other one is linguistics orientation approaches. The machine learning approach may be a supervised task because it involves the coaching of a classifier employing a assortment of representative information. On the opposite hand, the linguistics orientation approach involves the determination of the document's overall sentiment from the linguistics orientation of words it contains while not previous coaching and, thus, it's associate degree unsupervised technique. Chaovalit and Chow dynasty in 2005 compare the two same ways victimization reviews from the picture show domain[10]. The results show that the unsupervised linguistics orientation approach achieves low accuracy, however is far a lot of economical once employed in time period applications. In distinction, the supervised machine learning approach provides a lot of correct classification results however has the disadvantage that the coaching of the classifier tends to be terribly long. On account of this, researchers over and over apply unsupervised techniques so as to label a corpus that is later used for supervised learning.

A sentence level outline will give a deeper level of understanding of a subject. Marking the likelihood of every sentence to every topic victimization word likelihood in topic modeling of Topic Sentiment Mixture (TSM) model[11]. By selecting the highest stratified sentence in every class, they're able to show the foremost representative sentence. On the opposite hand, score sentences supported the TF-IDF of their words and choose the foremost relevant and discriminative sentence to be shown as summary[12]

Summary with a Timeline: Showed opinion trends over a timeline showed in[11,12]. General opinion summarization focuses on finding statistics of the ‘current’ knowledge. In reality, opinions modification as time goes by. Opinion outline with a timeline helps America see the trend of opinions a few target simply, and it can also tell America ideas for additional analysis.

To work out what changes people's opinions, we will analyse the events that happened at the forceful opinion modification. As an example, we will give the modification of opinions towards four election candidate, and that will simply determine that there's a forceful opinion modification on the day.

III. HEALTH FEATURE IDENTIFICATION USING LSA

In this paper, we tend to propose a unique approach supported LSA to spot health connected feature terms. Primarily, LSA could be a theory and technique to investigate relationships between a group of documents and therefore the terms they produce an ideas associated with the documents and terms. LSA could also be applied to any variety of count data over a definite domain that is alleged two-mode data[13]. Suppose we have a collection of documents say Docs= {d1, . . . , dn} with terms from Words = {w1, . . . , wm} are given, then the system can construct a term and document matrix M, where the dimension is n×m and each entry in the term and documents Mij denotes the number of times the term wj appeared in document di . Row vector and column vector is used to describe each document di and each term wi respectively.

Proposed Algorithm-I for Latent semantic analysis

```
Step 1 initializes an array F
Step 2 Convert term document matrixes and assign it as M
Step 3 Breaks the original matrix M into three compatible matrix using SVD
Step 4 for each supplied seed feature(S) find the semantically related feature f
        For $f \in S$ do
          $W_f \leftarrow$ TermVectorFromTermDocMatrix(f, $\hat{A}$)
          Initialize similarities list sml
                $i \leftarrow 1$
                for each column vector w of $\hat{A}$ do
                 $sml[i] \leftarrow W_f * W$
                      $I \leftarrow I + 1$
                End
                Sort (sml)
                Get toprelated features from $\hat{A}$
                End
Step 5 return F
```

Reduced singular-value decomposition (SVD) is the mathematical technique underlying a type of document retrieval and term similarity method known as latent semantic analysis and a low-rank approximation of the matrix M could be accustomed verify patterns inside the relationships between the terms and concepts contained inside the text

$$M = U\Sigma V^{T} \qquad (1)$$

U and V of equations (1) are matrices with orthonormal columns (i.e., $U^{T} U = V^{T} V = I$), and Σ is a diagonal matrix containing the square roots of Eigen values from U and V in descending order. The initial term-document matrix can enable the underlying latent relationships between phrases and documents to be exploited throughout searching. Equation (2) shows that the reduced matrix Â is obtained by reducing the dimensions k by preserving the important relationships. Therefore, in spite of the reality that the initial vector location is sparse, the corresponding low-dimensional place is generally no longer sparse. Much, the amount of dimensions preserved in LSA is associate empirical issue [14]. The different dimensions of experiments are conducted and explained the experiments section.

$$\hat{A} = U\Sigma' V^{T} \approx U\Sigma V T = M \qquad (2)$$

Singular value decomposition (SVD) is tested from three reciprocally compatible points of read. On the one hand, we will see it as a way for reworking related variables into a collection of unrelated ones that higher expose the assorted relationships among the initial knowledge things. At identical time, SVD is also a method for distinctive and ordering the dimensions thereon info points exhibit the foremost variation. This ties in to the third manner of viewing SVD, that is that after we've got known wherever the foremost variation is, it's potential to search out the simplest approximation of the initial knowledge points victimization fewer dimensions. Hence, SVD is wide accepted mathematically a method for knowledge reduction.

The basic concepts behind SVD is taking a high dimensional, extraordinarily variable set of points and reducing it to a lower dimensional house that exposes the substructure of the initial data further clearly and orders it from most variation to the tiniest quantity. What makes SVD smart for human language technology applications is that you can simply ignore variation below a selected threshold to massively deflate your data but be assured that the foremost relationships of interest area unit preserved.

The input to the Algorithm1 is a term-documents matrix, best seed features from health domain, the reduced spatial property in SVD operation, and then the range of available features can be extracted for every best seed features. We perform SVD operation on the term-document matrix to find the similarities between the seed product-feature vector and, pairwise, the opposite term vectors. The highest ones are going to be collected can be considered as related health feature for a specific health feature. The two procedure is employed, the first one is TermVectorFromTermDocMatrix is employed to get the term-vector illustration of a product feature. The best seed is meant to be one amongst the terms within the term-document matrix, and it's simple to get its corresponding document-vector illustration. Meanwhile, sml is employed to store the similarities between the best seed and therefore the different terms. Once it is sorting in descendant order, then we can get the highest ones and their corresponding feature names in procedure TopRelatedFeatures. Once these steps are completed, every health feature seed will have its own semantically connected term set. This technique can be applied to all or any the languages; it doesn't want any external word book, since LSA is language-independent, and it's supported algebra SVD operation.

IV. FEATURE BASED REVIEW SUMMARIZATION

Figure two shows design of projected Feature-based Reviews summarization system in which the input could be a doctor’s name and the most frequent feature is chosen to provide the user a short summary about the health service provider.

These reviews of doctors’ become the inputs of the SVM sentiment classifier, which could classify the reviews into high quality or poor classes. Score info is obtained based on the proportion of positive and negative health reviews. Moreover to the sentiment classification of fitness evaluation, we have a tendency to extra verify the polarity of a sentence victimization opinion words. Then, the device will supply every high quality and negative summarization, regardless of the polarity of a review sentences. The complete technique includes sentiment classification and characteristic-based summarization. These processes are delineated inside the following sections.

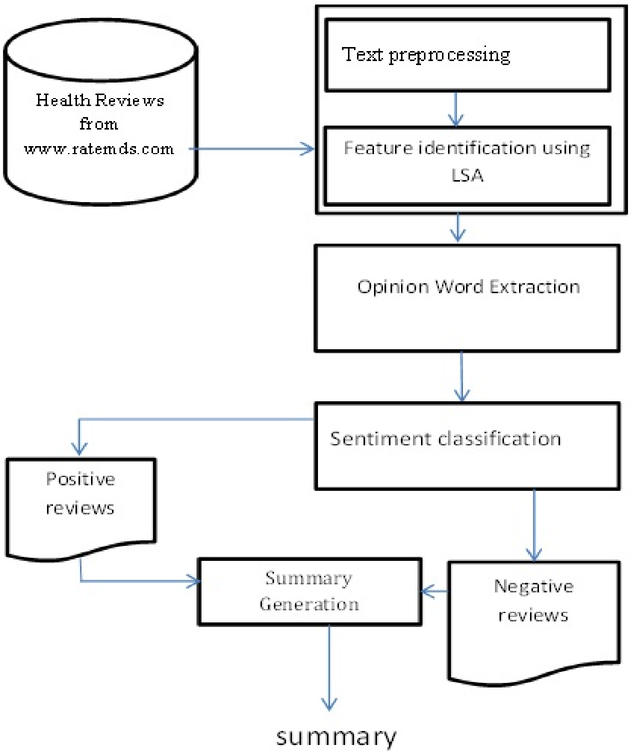


Figure 2. Architecture of Feature-based Reviews Summarization System

A. Dataset

We have collected the health reviews from www.ratemds.com. Since the first knowledge contains heap of extra info, we tend to performed clean up and pre-processing. It is necessary to provide the training data for SVM to build the classification model, and manually we split the training set into high-quality or low quality reviews. We've 480 positive and 480 negative reviews for building classification-model. Additionally to the model-building knowledge, we have collected around five thousands reviews of doctors from the internet.

B. Sentiment Classification

As mentioned on top of, sentiment classification is comparable to ancient binary-classification drawback .Currently, several classification algorithms like SVM[3,15,16,17], decision trees[18], and neural networks[19] are planned and shown their capabilities in numerous domains. SVM is one amongst the progressive algorithms. SVM measures the complexness of hypotheses supported the margin with that they separate the info rather than the amount of options.

In sophisticated NLP approaches, and data retrieval (IR), bag-of-words model tries to use unordered collection of words to represent a text, no matter the structure of language and its order. In alternative words, every word in the textual content contributes and also strengthens to the main feature of the document. We have decided to use similar approach to construct a feature vector of the document. Stop words are removed and stemming is employed to represent the root word then finding the distinct word can acts as feature. Using these features we can generate a feature vector, and machine-learning algorithms can be employed to perform classification tasks. We have employed SVM to perform the classification and libsvm[20] package is employed within the system. The kernel perform utilized in the system is that the radial basis perform (RBF) and K-fold cross validation is conducted within the experiment. Rating is calculated based on the proportion of positive and negative reviews; the system might give this information to the user. For instance if there are a hundred health reviews for a selected health supplier if eighty reviews are positive, the rating of this supplier are four stars.

C. Review Summarization

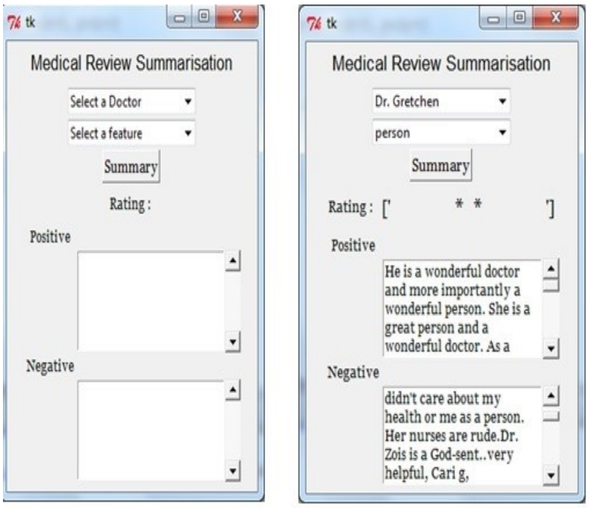
1) Health-Feature Identification: As mentioned higher as, we have a tendency to propose Associate in Nursing LSA-based health feature-identification formula and system will get a semantically connected feature set for every seed. We have a tendency to compare this feature-identification approach, i.e., the LSA-based approach, frequency-based approach.

2) Identification of Opinion-Word: Similarly to feature characteristic identification, extracting corresponding opinion words is more concerned for our summarization task. Hu and Liu extracted nearby adjective as the opinion words of product features [21]. Additionally to language sentence-structure characteristic, dependency graph is used to established relationship between feature words and the corresponding opinion words by Zhuang et al in training data[22].They each think about language syntax to extract opinion words; so, these approaches are applicable to those language sentences having such a properties.

Several languages don't possess the said syntax. Hence, we applied mathematical approach to get opinion words. First, we consider POS-tagging data of the opinion words. Second, term frequency is taken into consideration; so, frequency of the opinion words ought to exceed a threshold price.

3) Feature-Based Summarization: As represented on top of, feature-based report is a lot of acceptable in health review report system. In general, feature-based report is predicated on health options and opinion words. Hence, we have a tendency to propose associate degree LSA-based filtering approach to any

choose the content of the outline supported user's favour. In health feature discovery, we select an appropriate feature using LSA to search out health connected feature of a particular health service supplier, and these connected terms can be thought to be being semantically associated with this health feature. For every given health feature f, LSA may discover connected terms F that area unit semantically associated with f. In general F can be thought to be f's connected terms, and also the system will use F to pick out outline sentences. In application style, the system provides all the outline sentences within the starting. The health-feature seeds mentioned in LSA-based feature-identification method can become candidate interested options.



Dr. Gretchen

person

Rating : ['   * *   ']

He is a wonderful doctor and more importantly a wonderful person. She is a great person and a wonderful doctor. As a

didn't care about my health or me as a person. Her nurses are rude.Dr. Zois is a God-sent..very helpful, Cari g,

Figure 3. Summarization screen shot &Figure 4: Screen shots With reviews

The system permits the user to see the feature f within which he/she is interested. Once the user determines f, the system can generate an outline that is expounded to health options F. Much, a positive health review might embrace negative comments concerning specific aspects and contrariwise. During this paper, we have employed the sentiment polarity of a health review using SVM and obtained the sentence polarity using opinion words. In feature-based report, the system will utilize the polarity of opinion words to see the polarity of sentences. Hence, the system will give each positive- and negative-review report, in spite of the polarity of a review. The percentage of positive and negative reviews determines overall rating and reports it to the supplier or the heath consumer. The screen sections in Figure three and four permits the users in general and in particular to the health consumer to decide on the options within which they're interested. Whatever be the rating of service provider the system will provide the user all positive and negative sentences.

V. EXPERIMENT

Numerous experiments are done to validate our system. SVM is used to perform the sentiment-classification task and many feature combinations are accustomed value the system performance. When we run the application on mobile platform, the classification accuracy is not the sole criteria and but the factor response time is need to be considered. In health-feature identification, we tend to propose an LSA based mostly approach to spot the health options authored by health consumer and compare this approach with frequency-based.

A. Sentiment Classification

Opinions in linguistic communication are typically expressed in delicate and sophisticated ways in which. For instance, the polarity of a sentence is also modified once a negative term is employed within the sentence. Many experiments are performed to judge our system. SVM is used to perform the sentiment-classification task with many feature combination to validate the system performance. The systems response time plays very important role, therefore, system-response-time-evaluation experiment is conducted further. In product-feature identification, we tend to propose associate LSA based mostly approach to spot the health options and compare LSA-based approach with frequency-based approach.

We thought-about attainable feature combination within the experiments to get the most effective feature choice. Supported the bag-of-words model, we tend to used unigram, bigram, negation, location, frequency, and presence options (i.e., solely think about whether or not the feature is gift or not) to perform the classification task with completely different feature mixtures. Additionally to word segmentation, stop words are removed also, since the stop words cannot offer sufficient info. In feature choice, our experiments additionally showed that unigram with presence options outperforms written word with different options, and therefore the result's an as represented in table III. Additionally to unigram with presence options, we have performed experiments to under different conditions to match the variations of feature mixtures, and that they are represented as follows.

1) Type1:
a) Both positive and negative terms are removed from health reviews;
b) Criterion is chosen based on frequency
2) Type 2: frequency-feature criterion, where the term's square of frequency should be at least average and is defined as the average sum of square frequency.
3) Type3: The term should occur at least three times.

Table I shows the experimental result. We realized that the performance of type2 outperforms the type I and type III .Table II shows that the evaluation result of SVM classifiers on medical reviews and we received the value of precision, recall and F1-score can be compared to other standard approach. Unigram with presence feature can have 39000 options, and it takes concerning a hundred and fifteen s to load the classification model. Obviously, it's unfeasible on mobile platform if a system's response takes a hundred and fifteens. Table III shows that it takes nearly six second to load classification model, and it's possible on mobile platform. Therefore, this frequency criterion is used to perform sentiment classification.

Table 1. Experimental Result of Various Feature Combinations

Features	**Accuracy**
Unigram with presence of feature	82.20%
Type I	68.00%
Type I + negation	67.00%
Type II	75.06%
Type II + negation	76,32%
Type II + position	68.04
Type III	73.05%
Type III+ negation	72.08
Type III + position	67.15

Table 2. Svm Classifiers Experiments With Reviews On Doctors

Classifier	Sentiment	Precision	recall	F1-score
SVM	-1	0.60	0.32	0.41
	0	0.69	0.90	0.78
	1	0.65	0.40	0.50

Table 3. Svm Model Loading And Prediction Evaluation Result(Sec)

Feature type	**No of features**	**Model loading**	**prediction**
Frequency-based	85	.25	< .0625
Unigram with presence	39000	115	0.5-0.625

B. Health-Feature Identification

In health-feature identification, we tend to compared our LSA based mostly approach with different approach, that is frequency based mostly. Precision, recall, and F-value area unit used to judge system performance. In frequency-based approach, all the nouns area unit hierarchic in line with their frequencies, and then, the highest ones area unit elite as health options. Table IV shows the highest 10 terms using frequency-based approach.

Table 4. Top Ten Health Features Identified Using Frequent Features

Time
Recommend
Care
Appoint
Health
Online
Concern
Wonder
Friendly
Kind
Treatment

Hence, the terms like decision, results, blood and take a look at are often known. Within the LSA-based approach, Algorithm1 is employed to spot product options and also the seeds embrace Prescription & Tests, Attention and recommendation. The truncated dimension of LSA is five hundred during this paper. Table V shows the highest four options for every seed. Additionally to health-feature identification, the highest four options for every seed are often thought to be being semantically associated with the seed.

C. Discussion

In sentiment classification, Pang Showed that unigram with presence options outperformed different feature combinations3. Our experiments adapt to Pang's analysis results. However, if the entire unigrams area unit utilized in the system, the amount of options are going to be huge. As an example, our coaching dataset includes 980 health reviews, and therefore the variety of options is around 39,000. The appliance must load SVM model 1st and so predict the linguistics orientation of the review. In health-feature identification, the experiment shows that LSA-based approach outperforms frequency-based. Our LSA-based system will establish semantically connected term set for every seed. We have a tendency to propose associate degree LSA based mostly filtering mechanism to use these semantically connected terms to scale back the scale of outline. Solely the sentences containing these terms are going to be conferred to users. Moreover, the LSA based identification approach may be generalized to different service domains, since SVD operation may be carried out to any language. However, the health-review dataset doesn't possess such a characteristic. The articles within the health review area unit similar, since all of them target health reviews. Presently, feature-based document is sentence-level report. Though the sentences are extracted totally from different paragraphs or health reviews, therefore it's difficult to have high fluency in the summary. Thus, in future work we will overcome the shortcoming.

VI. CONCLUSION

In this paper, we've planned feature primarily based health review-summarization system. Rating info is obtained from the sentiment-classification results. In this report, LSA is used for health-feature identification, and it extracts all the connected health features. Moreover, we have a tendency to use a mathematical approach to spot opinion words. Salient health options and opinion words are used as the basis for feature-based report. The quantity of options plays a very important role in SVM-model loading and prediction. We have used frequency criterion based options, and therefore the experiment shows that it takes but nearly six second to load the SVM model and classify the reviews. The planned approach during this paper may absolutely utilize the net content to produce a replacement health-review report and rating service. The planned designed can even be extended to different domains simply.

VII. REFERENCES

[1] Mani I, Klein G, House D, Hirschman L, Firmin T, Sundheim B. SUMMAC: A text

summarization evaluation. Natural Language Engineering. 2002;8(01):43–68.

[2] Afantenos S, Karkaletsis V, and Stamatopoulos P. Summarization from medical documents: a survey. Artificial intelligencein medicine. 2005;33(2):157–77.

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

[4] Turney P. D, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in Proc. 40th Annual Meeting Assoc. Comput. Linguist., 2002, pp. 417–424.

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

[6] Choi S.H, Jeong Y. -S, and Jeong M. K, "A hybrid recommendation method with reduced data for large-scale application," IEEE Trans. On Syst.,Man, and Cybernetics. C: Appl. Rev.2010 sep; vol. 40, no. 5, pp. 557–566

[7] Mullen T and Collier N, "Sentiment analysis using support vector machines with diverse information sources," in Proc. EMNLP. 2004; pp. 412–418.

[8] 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.

[9] 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.

[10] Chaovalit P., & Zhou, L. (2005). Movie review mining: A comparison between supervised and unsupervised classification approaches. In Proceedings of the 38th Hawaii international conference on system sciences, 112.3.

[11] 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

[12] 12. 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

[13] Hofmann T, Puzicha J, and Jordan M.I, "Learning from dyadic data," in Proc. Conf. Adv. Neural Inform. Process. Syst. II, Cambridge, MA: MIT Press, 1999, pp. 466–472.

[14] Landauer T K, Foltz P W, and Laham D, "Introduction to latent semantic analysis," Discourse Processes, vol. 25, pp. 259–284, 1998.

[15] Vapnik V. N.,The Nature of Statistical Learning Theory. NewYork:Springer-Verlag, 1995

[16] Joachims T, Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. Norwell, MA: Kluwer, 2002.

[17] Silva C, Lotri ˇc C, Ribeiro B, and Dobnikar A, "Distributed text classification with an ensemble kernel-based learning approach," IEEE Transactions on .System, Man. and Cybernetic. C: Appl. Rev., vol. 40, no. 3, pp. 287–297, May 2010.

[18] Rokach L and Maimon O, "Top-down induction of decision trees classifiers—A survey," IEEE Trans. Syst., Man, Cybernetic. C, Appl. Rev.,Vol. 35, no. 4, pp. 476–487, Nov. 2005.

[19] Zhang G.P, "Neural networks for classification: A survey," IEEE Trans.Syst., Man, Cybernetic- C, Appl. Rev., vol. 30, no. 4, pp. 451–462, Nov. 2000.

[20] (2001). LIBSVM: A library for support vector machines [Online].Available:http://www.csie.ntu.edu.tw/ cjlin/libsvm.

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

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