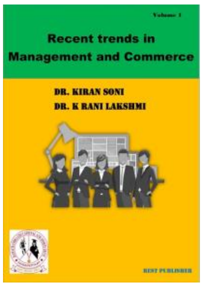
Recent trends in Management and Commerce
Vol: 1(1), 2020
REST Publisher
ISBN: 978-81-936097-6-7

Website: http://restpublisher.com/book-series/rmc/



Predicting HCAHPS scores from hospital reviews and social media pages

Prof. Mrs.Sujata Khedkar[1], Smith Gajjar[2], Hrithik Malvani[3], Jai Soneji[4], Sonia Thakur[5].
[1]Associate Professor, Department of Computer Engineering, VESIT, University of Mumbai, India.
[2,3,4,5] Department of Computer Engineering, VESIT, University of Mumbai, India.

Abstract

Nowadays, we can find any information related to any business firm or any facilities easily on the internet. But sometimes the available data is not in the required format and may need some processing. Once the processing of data is done it can be used for various purposes. Similarly, we can find many hospital websites on the internet and we can also read the reviews given by patients who had already visited that particular hospital. But this data is available in different forms and at different places. In this paper, we focus on the problem of predicting HCAHPS scores from hospital reviews and social media pages. Some existing examples of HCAHPS parameters include communication with doctors and nurses, the responsiveness of hospital staff, the quietness and cleanliness of the entire hospital environment, relevance of medicines, discharge information and overall rating of hospital. The data is first collected from different sources, which is then processed and applied to different algorithms. Proper prediction of the HCAHPS score of the hospital will help people to understand and go for better treatment.

Keywords: HCAHPS - Hospital Consumer Assessment of Healthcare Providers and Systems, FS - Feature sets,

1. Introduction

Customer/Google Reviews are extremely useful to gather information about the working of any organization/hospital. We can gather information about a particular organization/hospital from various sources like the internet or face to face communication. When people need to visit a hospital they try asking their relatives or their friends to get information about a hospital. But when there is some urgency and if none of the relatives or friends is available or if a person is a new resident of a particular area and if they are in search of a good hospital then they can take help from the reviews given by the other people. So here social media plays a very important role in helping these people to find a place for better treatment. People who have visited the hospital also play a very important role by writing reviews for a particular hospital from their experiences. The information available on the internet can be helpful to many people and can save their time in finding a good place for their treatment. As reviews posted by people may contain various kinds of information, it might be useful to automatically identify the exact nature of the information that is present in a given review. A review posted on the internet can contain much useful information like the nature of doctors, cleanliness, the infrastructure of the hospital, etc. In many cases, a single review may contain information about multiple categories. This review can be termed as a multi-class single-label classification problem and need to run different algorithms to solve this type of classification problem. In some reviews, the reviewer writes very short reviews for eg: “good” or some time reviewer writes a review in an informal way for eg: “use of smileys or abbreviations” this acts as a major problem while classifying the reviews. The contributions of the work are given below: – We identify different feature sets for representing the reviews. Along with tf–idf features we use a few features derived from the tweet collection. The performance of each classifier, for different feature sets, is analyzed in detail. We also evaluate the effect of adding extra features in detail. The structure of the rest of this paper is as follows. We discuss related work from literature in Sect. 2. We further discuss the use of social media for predicting HCAHPS scores from the dataset containing information in Sect. 3. Then, in Sect. 4, we define the problem of tweet classification. A description of the classifiers and various features used in the work are presented in Sect. 5. Our experimental set up is discussed in 6. Experimental results are presented and discussed in Sect. 7. We conclude the paper with a brief discussion of our findings in Sect. 8

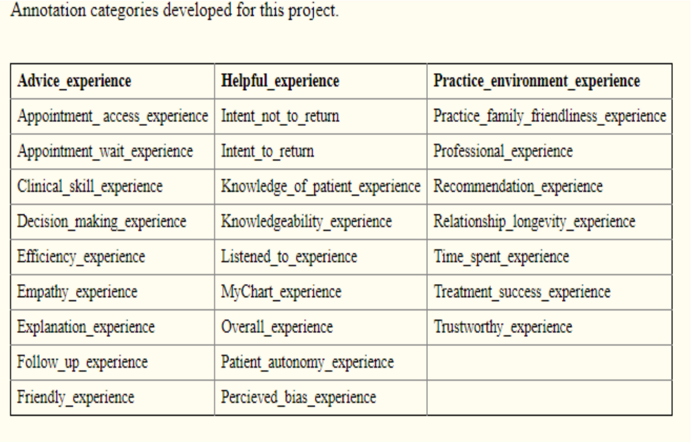
2. Related work:

In this section, different ways of collecting textual data (Customer reviews) has been explored and also how scores are given on NLP Projects

Analyzing Free-text comments: In [1], the author determines the topic from the customer textual reviews to know their context and also the author has implemented automatic topic classifiers and worked on finding the negativity of comments by using Sentiment analysis and for detailed topics, the author has determined the common topics within the negative comments. After sentiment analysis, a total of 28 topics were determined but only 7 most frequent were considered. For automated topic classification, they developed vocabulary-based and Naive Bayes classifiers. The free-text comment fields, which are filled out in nearly 50% of patient surveys, are underutilized. The Center for Medicare and Medicaid Services

(CMS) and Agency for Healthcare Research and Quality (AHRQ) developed a national standard for reporting patient satisfaction called the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS).

Annotation Study: In [1], the author developed a model for characterizing topics from patient survey responses (Figure 1) and validated our schema with an annotation study.



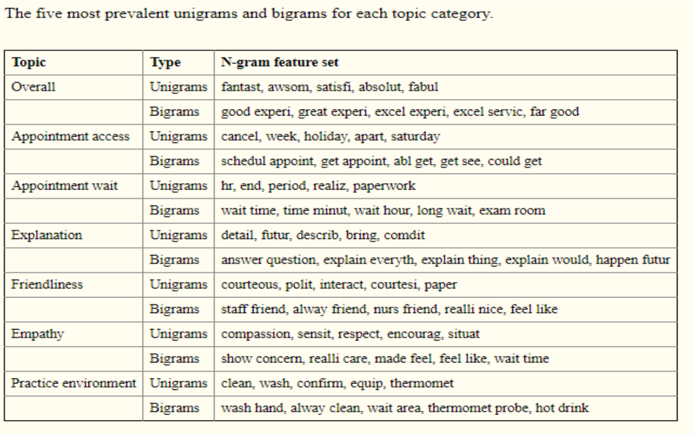
Annotation categories developed for this project.

Advice_experience	Helpful_experience	Practice_environment_experience
Appointment_ access_experience	Intent_not_to_return	Practice_family_friendliness_experience
Appointment_wait_experience	Intent_to_return	Professional_experience
Clinical_skill_experience	Knowledge_of_patient_experience	Recommendation_experience
Decision_making_experience	Knowledgeability_experience	Relationship_longevity_experience
Efficiency_experience	Listened_to_experience	Time_spent_experience
Empathy_experience	MyChart_experience	Treatment_success_experience
Explanation_experience	Overall_experience	Trustworthy_experience
Follow_up_experience	Patient_autonomy_experience	
Friendly_experience	Percieved_bias_experience	

Figure 1

They developed an annotation schema for topics and recorded sentiment in free-text satisfaction responses. They added and removed categories in consultation with the Exceptional Patient Experience team (author CD) from UUHSC. In total, 28 annotation categories were listed. Words were annotated with appropriate categories. They also took sentiment into consideration viz. Positive, negative and Neutral. 63% of total annotations were chosen from 7 common topics viz. overall, appointment access, appointment wait, explanation, friendliness, practice environment, and empathy.

Topic Classification: In [1], the author developed both vocabulary-based and machine learning-based approaches. For the vocabulary-based method, they used the annotated topics in the 300 adjudicated documents to generate a vocabulary. Then they gathered the text from each topic and used the Natural Language Toolkit (NLTK) for Python to tokenize the comments, remove non-alphabetic characters, stem (Snowball Stemmer) the resulting tokens, and remove stop words from each tokenized comment. The results were converted into n-grams: unigrams, e.g., “bad,” and bigrams, e.g., “bad environment”. The lists of n-grams were compared so that each n-gram appeared on only one list. If an n-gram was on more than one list, the list with the highest frequency of its occurrence got to retain it. The top five n-grams for each topic are listed in Figure 2. From Figure 2, they observed that the top n-grams associated with appointment access and appointment wait are associated with time. In contrast, the practice environment and n-grams are often associated with cleanliness and temperature. Empathy and friendliness n-grams describe feelings and service actions.



The five most prevalent unigrams and bigrams for each topic category.

Topic	Type	N-gram feature set
Overall	Unigrams	fantast, awsom, satisfi, absolut, fabul
	Bigrams	good experi, great experi, excel experi, excel servic, far good
Appointment access	Unigrams	cancel, week, holiday, apart, saturday
	Bigrams	schedul appoint, get appoint, abl get, get see, could get
Appointment wait	Unigrams	hr, end, period, realiz, paperwork
	Bigrams	wait time, time minut, wait hour, long wait, exam room
Explanation	Unigrams	detail, futur, describ, bring, comdit
	Bigrams	answer question, explain everyth, explain thing, explain would, happen futur
Friendliness	Unigrams	courteous, polit, interact, courtesi, paper
	Bigrams	staff friend, alway friend, nurs friend, realli nice, feel like
Empathy	Unigrams	compassion, sensit, respect, encourag, situat
	Bigrams	show concern, realli care, made feel, feel like, wait time
Practice environment	Unigrams	clean, wash, confirm, equip, thermomet
	Bigrams	wash hand, alway clean, wait area, thermomet probe, hot drink

Figure 2

They compared two methods for classifying comments into each topic. First, they used a simple dictionary lookup approach. For each unigram and bigram in a comment, they found the corresponding topic. All matched topics were retained because many comments had more than one hand annotation. Like all attempts at document classification, there was a tradeoff between identifying the document category correctly (i.e., precision) and finding all the documents that belonged in that category (i.e., recall). They observed that allowing n-grams to appear in the lists of more than one topic increased recall at the cost of precision. For comparison against the vocabulary-based approach, they trained a machine learning approach leveraging the Naive Bayes (NB) algorithm. They tested decision tree and SVM models as well, but they found the best performance with Naive Bayes’. Then they used the NLTK (i.e., featx) methods to reduce the text to lowercase, stem words to their lemma, group the stemmed words into n-grams, and convert the n-grams into feature vectors. They used the NLTK NB classifier, with default settings. They divided the data set into training/test data sets (75%/25%), one binarized set per annotated topic, each set was balanced between "topic containing" and "no topic" comments (i.e., positive and negative examples of the class). They applied the trained algorithm to the blind test set.

Sentiment analysis: In [1], the author used the same training and test sets and encoded n-gram features from the topic classification, we developed a classifier to categorize a comment based on its sentiment. They trained the NLTK NB approach to classify comments based on sentiment categories of positive or negative. They reported the performance for each sentiment category on the test set. Then they ran both the trained and tested NB topic classifier and NB sentiment classifier on the remaining roughly 50,000 patient satisfaction comments. They reported the distribution of positive and negative comments by topic class.

Topic Modeling: In [1], in order to complement the topics annotated through manual review, the author completed an unsupervised topic modelling study. They first classified the full 51,234 comments with sentiment categories using the NB sentiment classifier. For all comments classified as negative, they provided the unigrams and bigrams to an LDA algorithm with a preset maximum of 30 topics using the gensim package. They reported the n-grams associated with 10 of the 30 topics learned by the algorithm and if a topic suggests one of our seven most common topics, they also provide a topic label.

Tf–idf features: In [2], the author used tf-idf for representing the tweets in their work. Tf–idf stands for Term Frequency—Inverse Document Frequency. After the addition of bigrams, the number of features almost increased by threefold. They used the tf–idf features unigram and bigram as two base feature sets and created 8 feature sets by using different combinations of derived features and manual features. Each of the combinations is described below:

– FS1 (unigrams): This feature set consists of tf–idf values of unigrams only.
– FS2 (unigrams 1 bigrams): This feature set contains tf–idf values of unigrams along with bigrams.
– FS3 (unigrams 1 manual): To create this feature set they combined unigrams FS1 with manually created features.
– FS4 (unigrams 1 bigrams 1 manual): In this set, they added unigrams and bigrams FS2 with manual features.
– FS5 (unigrams 1 derived): Combination of unigrams FS1 with derived features create this set.
– FS6 (unigrams 1 bigrams 1 derived): They combined unigrams and bigrams FS2 with derived features to create this set.
– FS7 (unigrams 1 derived 1 manual): They added derived features with FS3 for this set.
– FS8 (unigrams 1 bigrams 1 derived 1 manual): Derived features along with FS4 creates this set.

3. Problem definition

The internet has a huge number of customer reviews which express different opinions and concerns, one has to go through all that to make a credible decision on anything, our prototype collects all those textual reviews using website content scrappers and performs different algorithms to give a score to the Hospital. It would help hospital management to check their score and would help to improve their services. The algorithm classifies a review into one of HCAHPS Parameters. These Parameters determine the final score of a Hospital. The concept we have used is Natural Language Processing and has been implemented with different Machine Learning Algorithms. This has been divided into different phases starting from Data collection followed by Data preprocessing, Sentiment Analysis, Topic modelling, Classification and finally a Scoring algorithm.

4. Methodology:

We are modelling this problem as a multi-class single-label unsupervised classification problem, where we can predict a single class label for each instance. In our model, we have first performed topic modelling on the review text to model customer reviews for categorizing into different topics and by using this modelled data which contains reviews text, as well as the dominant topic, is then fed to the Classification Algorithm. For topic modelling, we have used the LDA Topic Modelling Algorithm.

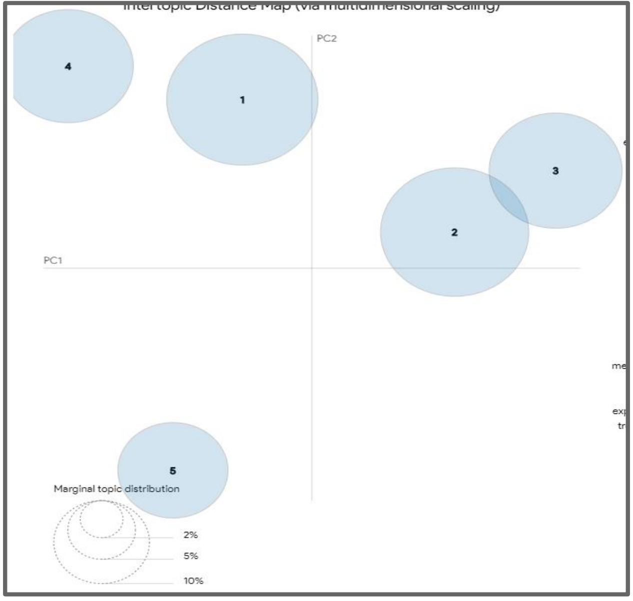
Topic Modelling Algorithm

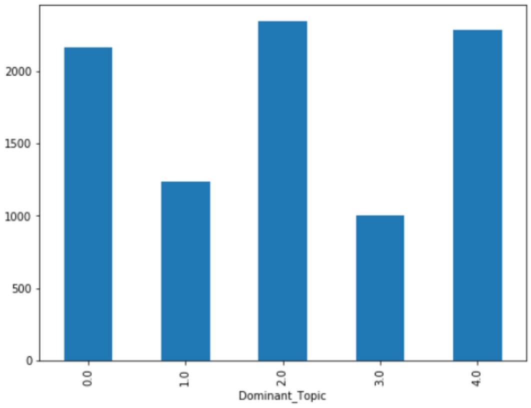
Id2word: In Id2word, we are assigning a unique integer id to all words appearing in the corpus with the gensim. corpora. Dictionary. Dictionary class. This will sweep across the texts, collecting word counts and the relevant statistics. The function doc2bow() will simply count the number of occurrences of each distinct word, convert the word to its integer word id and then will return the result as a sparse vector.

Tf-idf: We first compute the TF-IDF. In order to do this we are multiplying a local component and a global component then we are normalizing the resulting documents into a unit length. The formula for non-normalized weight of term i in document j in a corpus of D documents:

$$weight_{i,j} = frequency_{i,j} * log_2 \frac{D}{document\_freq_i}$$

Latent Dirichlet allocation: We are using Latent Dirichlet Allocation as it explains some unobserved groups which form a set of observations which in turn explains why some parts of the data are similar. LDA is a "generative probabilistic model" of a collection of composites made up of parts. It uses Natural Language Processing (NLP) and topic modelling, among with others. In terms of topic modelling, the composites which are documents and the parts which are words or phrases (phrases 'n' words in length are referred to as n-grams). To enhance this accuracy of LDA we have implemented statistical measure tf-idf. After the implementation of LDA with 5 topics, we visualized the results with pyLDAvis. Then the LDA classified each review text into multiple topics, we then took the average of all the weights assigned to each word and we decided the dominant topic for that review text. As we are using a Multi-class single-label classification, we had to classify only one dominant topic to a particular review text. Later on, after performing topic modelling we fed this data to the Machine learning algorithms in order to create a model. For training this model, we have used multiple algorithms to get the highest accuracy.





Learning Algorithms:

Random Forest: Random forest is an ensemble approach that consists of multiple decision trees each trained on a sub-sample of the data. In addition to this introducing randomness in the training data, it can also add randomness in feature selection. Instead of selecting the best feature among all the features for a split at a node, it only selects the best feature among the random subset of features. Random forest is found to achieve better generalization performance by avoiding over-fitting through model averaging. Prediction is done by choosing a class with the majority of the voting from individual decision trees.

Linear SVC: The linear-SVM which includes a linear kernel for its basis functioning. It is much less tunable and is basically just a linear interpolation. The objective of a Linear SVC (Support Vector Classifier) is to fit the data we have provided and then returning a "best fit" hyperplane which either divides or categorizes our data. From here, after getting the hyperplane, we then feed some features to our classifier to see what the "predicted" class is. It returned an accuracy of 0.44.

Logistic Regression: Logistic regression is a statistical machine learning algorithm which can classify the data by considering the outcome variables on extreme ends and then tries to make a logarithmic line that distinguishes between them.

Formula: y=mx+c

where, y = value that has to be predicted; m = slope of the line; x = input data; c = y-intercept

This algorithm gave an accuracy of 0.457143

MultinomialNB: As in Multinomial Naive Bayes, for a class the probability of any particular word is estimated, as the relative frequency of any term t in documents belonging to the class. These variations are taken into account as the number of occurrences of term t in the training documents from class, which will include multiple occurrences.

5. Experimental setup:

During the implementation, we divided this information into 2 sets with 70% and 30% of the entire data.

Data preprocessing: Before working with this data, we first cleaned the data by performing the below-mentioned steps in sequence.

Stop-words and punctuation removal: We removed each and every word present in the NLTK stopwords.

Stemming: We removed the suffix from a word and reduced it to its root word.

Sentiment Analysis: By using sentiment analysis, we calculated the polarity of a text review and then removed the text review which has a polarity of 0 because it is not contributing for analysis as it doesn't give any opinion.

Named Entity Recognition:

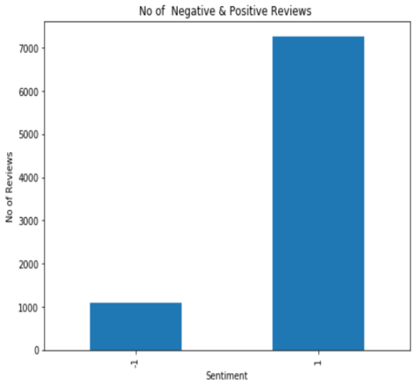
We used NER to remove Nouns and non-English words.

Metrics Used:

We used precision, confusion-matrix, recall to evaluate the performance of the algorithms. The definition of the above-mentioned algorithms are given below.

Precision: For a given class, out of all the text reviews that are assigned with the dominant topic, what fraction of them actually belong to that dominant topic, is called the precision. If Actual denotes the set of text reviews which belongs to that topic and Predicted denotes the set of Text reviews which are assigned to that topic by the algorithm.

$$Precision = \frac{|Actual \cap Predicted|}{|Predicted|}$$



Recall: For a given class, out of all the Text reviews that are from that topic, what fraction of them are predicted to belong to that same topic is called the Recall. If Actual denotes the set of Text reviews that belong to that topic and Predicted denotes the set of Text reviews that are assigned to that topic by the algorithm.

$$Recall = \frac{|Actual \cap Predicted|}{|Actual|}$$

Topics	precision	recall
1.0	0.77	0.81
2.0	0.59	0.59
3.0	0.78	0.82
4.0	0.77	0.58
5.0	0.78	0.77

F1 score: F1 score for a class is the harmonic mean of Precision and Recall obtained for the class. F-measure for the class can be computed as:

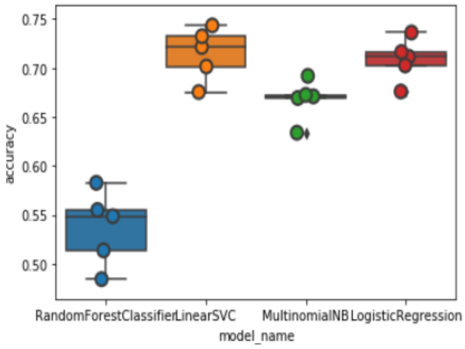
$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Confusion-matrix: We used a confusion matrix which is a summary of prediction results on the basis of the classification problem. The number of correct and incorrect predictions are then summarized with the count values and then is broken down by each class. This is the key to the confusion matrix. The confusion matrix which shows the ways in our classification model is confused when it makes predictions. It gives us the insight of not only into the errors being made by the classifier but also the types of errors that are being made.

```
[[586,  25,  61,  17,  36],
 [ 29, 227,  38,  12,  76],
 [ 78,  30, 658,  12,  25],
 [ 32,  24,  43, 181,  31],
 [ 39,  76,  42,  14, 588]]
```

6. Result

This section describes results obtained from previously considered feature sets and different algorithms. We used a tf-idf feature consisting of unigrams + bigrams) and implemented 4 different classification algorithms, namely RandomForestClassifier, LogisticRegression, LinearSVC, MultinomialNB. For classification, we used algorithms with the highest accuracy. We have used a boxplot graph for visualizing the accuracies of algorithms.



We got different results as LinearSVC with **0.737421** accuracy ,LogisticRegression with 0.732159 accuracy ,MultinomialNB with 0.6678 accuracy ,RandomForestClassifier with 0.5370 accuracy with Feature Set. We calculated the accuracy of all classifiers with FS1, FS2, FS3.

Classifiers	FS1	FS2	FS3

LinearSVC	**0.737421**	0.717094	0.712311
Logistic Regression	0.732159	0.711592	0.708603
MultinomialNB	0.695687	0.668424	0.661968
Random ForestClassifier	0.546106	0.53546	0.528528

Among all the algorithms LinearSVC got the highest accuracy of 0.7374 with a feature set 1 (unigrams). So we considered the LinearSVC algorithm with FS1. We carried out f1-score for all the topics using LinearSVC algorithm on all the feature sets.

F1-Score	
Topics	**Score**
Staff	0.78
Doctor	0.78
Cleanliness	0.79
Infrastructure	0.65
Discharge Info	0.76

Further, we developed a scoring model for all individual classes and also the overall score for HCAHPS. A normalised scoring model was needed due to unbalanced data present in each class. Scoring was given in range from 0 to 10. We scored every individual class within the range and also calculated the overall HCAHPS score for Apollo Hospital Hyderabad.

HCAHPS SCORE	
Topics	**Score**
Staff	7
Doctor	6
Cleanliness	7
Infrastructure	5
Discharge Info	7
Overall HCAHPS Score	**6**

7. Conclusion and future work:

In this paper, we discussed our prototype, which is one of the services that will be able to rate Hospitals & visualize different parameters of the hospital on the basis of HCAHPS parameters. The reviews are gathered from Google reviews and other social sites like practo. We have implemented the data pre-processing model in which the stopwords are removed, tokenized. We also implemented the LDA model which takes input as the bow corpus, dictionaries and the number of topics and mapped the topics to the original documents. The output of the data pre-processing model is given as input to the sentiment Analysis model, where the sentiment analyser categories the words as positive and negative words. We implemented the Multi class single label classifier. In future we plan to further improve our LDA and classifier model along with frontend (visualization) as the project is implemented as a service. Implement Graphical User Interface(GUI) to visualize and display results that are easily understandable by the end-users. As this service is for only one hospital, in future this can be made for multiple hospitals. The service can also be made from Realtime comments from various sources.

References

[1]. Kristina Doing-Harris, PhD,1 Danielle L. Mowery, PhD,3 Chrissy Daniels, MS,2 Wendy W. Chapman, PhD,3 and Mike Conway, PhD3: Understanding patient satisfaction with received healthcare services: A natural language processing approach.

[2]. Samujjwal Ghosh, P. K. Srijith, Maunendra Sankar Desarkar: Using social media for classifying actionable insights in disaster scenarios.

[3]. Archana B. Salunke, Smita L. Kasar "Personalized Recommendation System for Medical Assistance using Hybrid Filtering",, Jalgaon Road, Aurangabad, India, Oct 2015.

[4].Mohammad Reza Khoie, Tannaz Sattari Tabrizi, Elham Sahebkar khorasani Shahram Rahimi and Nina Marhamati, "A Hospital Recommendation System Based on Patient Satisfaction Survey", www.mdpi.com, 21 September 2017

[5].Sudarshan S., Kayathi Rohith, K. P. Sai Krishna, M. V. Panduranga Rao, " AutoHS: The Intelligent Hospital Search ", Indian Institute of Technology Hyderabad, India, 2014.

[6].James I. Merlino, MD, FACS, FASCRS, Carmen Kestranek, Daniel Bokar, Zhiyuan Sun, MS, Steven E. Nissen, MD MACC, David L. Longworth, "HCAHPS Survey Results: Impact of Severity of Illness on Hospitals " Performance on HCAHPS Survey Results " MD, FACP, November 1, 2014

[7]. Ajinkya Kunjir, Jugal Shah, Navdeep Singh, Tejas Wadiwala, "Big Data Analytics and Visualization for Hospital Recommendation using HCAHPS Standardized Patient Survey", Lakehead University, Department of Computer Science, Thunder Bay, Ontario, Canada. 2019

[8]. Masumi Okuda, Akira Yasuda, Shusaku Tsumoto, "A Data Mining Approach on the Structure of Patient Satisfaction in HCAHPS Databases".

[9]. Libin Zhang, Wei Wang, "Learning to Rank with Bayesian Evidence Framework ", International Conference on Computer Science and Software Engineering, 2018

[10].Tyler Vovos, M.D., Sean Ryan, M.D., Cierra Hong, B.S., Claire Howell, B.S., Thomas Risoli, Jr., M.S., David Attarian, M.D., Thorsten Seyler, "Predicting inpatient dissatisfaction following total joint arthroplasty: An analysis of 3,593 HCAHPS survey responses", M.D., PhD, Investigation performed at Duke University Medical Center, Department of Orthopaedic Surgery, Durham.

[11]. Giordano, Laura & Elliott, Marc & Goldstein, Elizabeth & Lehrman, William & Spencer, Patrice. (2009). " Development, Implementation, and Public Reporting of the HCAHPS Survey ". Medical care research and review.

[12]. Elliott, Marc & Lehrman, William & Goldstein, Elizabeth & Hambarsoomian, Katrin & Beckett, Megan & Giordano, Laura. (2009). " Do Hospitals Rank Differently on HCAHPS for Different Patient Subgroups? ".

[13]. Tabrizi, Tannaz & Khoie, Mohammad Reza & Sahebkar, Elham & Marhamati, Nina. (2016). Towards a patient satisfaction based hospital recommendation system

Copyright@ REST Publisher