Journal of Universal Computer Science, vol. 22, no. 5 (2016), 650-670
submitted: 15/9/15, accepted: 29/4/16, appeared: 1/5/16 © J.UCS

Feature Based Sentiment Analysis for Service Reviews

Ariyur Mahadevan Abirami
(Thiagarajar College of Engineering, Madurai, India
abiramiam@tce.edu, abiramiam77@gmail.com)

Abdulkhader Askarunisa
(Vickram College of Engineering, Madurai, India
askarunisa@staff.vickramce.org, nishanazer@yahoo.com)

Abstract: Sentiment Analysis deals with the analysis of emotions, opinions and facts in the sentences which are expressed by the people. It allows us to track attitudes and feelings of the people by analyzing blogs, comments, reviews and tweets about all the aspects. The development of Internet has strong influence in all types of industries like tourism, healthcare and any business. The availability of Internet has changed the way of accessing the information and sharing their experience among users. Social media provide this information and these comments are trusted by other users. This paper recognizes the use and impact of social media on healthcare industry by analyzing the users' feelings expressed in the form of free text, thereby gives the quality indicators of services or features related with them. In this paper, a sentiment classifier model using improved Term Frequency Inverse Document Frequency (TF-IDF) method and linear regression model has been proposed to classify online reviews, tweets or customer feedback for various features. The model involves the process of gathering online user reviews about hospitals and analyzes those reviews in terms of sentiments expressed. Information Extraction process filters irrelevant reviews, extracts sentimental words of features identified and quantifies the sentiment of features using sentiment dictionary. Emotionally expressed positive or negative words are assigned weights using the classification prescribed in the dictionary. The sentiment analysis on tweets/reviews is done for various features using Natural Language Processing (NLP) and Information Retrieval (IR) techniques. The proposed linear regression model using the senti-score predicts the star rating of the feature of service. The statistical results show that improved TF-IDF method gives better accuracy when compared with TF and TF-IDF methods, used for representing the text. The senti-score obtained as a result of text analysis (user feedback) on features gives not only the opinion summarization but also the comparative results on various features of different competitors. This information can be used by business to focus on the low scored features so as to improve their business and ensure a very high level of user satisfaction.

Keywords: Sentiment analysis, Opinion mining, Sentiment classifier, TF-IDF, Linear regression, online reviews
Categories: I.2.7, H.3.3

1 Introduction

The Internet world provides a number of review sites today. The amount of reviews in social media websites is getting increased day by day. Many people feel that they can come to a decision based on the suggestions and experiences of other people by analyzing reviews given by them. This type of analysis on user reviews or customer feedback has become important now-a-days. Enterprises, Corporate, Service providers do social media analytics to improve their business, customer relationship, etc. People start using online health services to get social support through social interactions. Understanding the emotional impact of online users enables to improve the facilitation for other social network members. The power of social media is described by [JagadeeshKumar, 14] and it can be effectively used to influence public opinion or research behavior.

Most of the reviews are written in unstructured or semi-structured format. People express their feelings or share their experiences through these blogs, review sites, etc. All these unstructured text and voluminous data about sentiments have to be analyzed manually or semi-automatically to understand the context and to make decisions or improvements in the business. A single sentence may contain multiple opinions, subjective and factual clauses. Knowing the polarity of a particular feature of a product or service is more useful than knowing the opinion of the sentence whether it is positive or negative. In particular, feature based product review summarization helps reader understand which feature is being highlighted in the review. User review analysis has become an active research topic in the past few years, as they can be seen as a major control mechanism for decision making. User shares his opinion with the overall topics of interests and also he can make use of star ratings. The measure of success relies on the analysis of the text reviews along with the star ratings. In recent years, it has been witnessed that opinionated postings in social media have helped in business improvement, and have great impact on social and political systems [Liu, 12]. Thus it becomes vital to collect and study opinions from the users. Figure 1 explains the generic framework for opinion mining and opinion summarization.

Figure 1: Sentiment Analysis Framework

Information Extraction (IE) from the online review comments becomes essential in order to make automatic decisions, to improve business intelligence, etc. Sentiment Analysis is one of the applications of IE. Sentiment analysis is defined as the automatic extraction of subjective content from digital text and predicting the subjectivity as positive or negative. Sentiment analysis or opinion mining is the computational study of people's opinions, sentiments, evaluations, attitudes and emotions expressed in written text [Liu, 12]. Sentiment analysis applications analyze feeds from social media data such as blogs, review sites, twitter, Google reviews, etc. and help business get deep insights into how a review can be categorized into variety of features, a topic of interest and so on. This is achieved by extracting words from the reviews posted in the websites that refer to positive or negative sentiments about a particular feature. The model proposed in this paper identifies the feature and its related opinion pair words. For example, a review says that “Very good machines and equipments which though considering the cost of hospital sound normal, but still no satisfaction for me because I had experienced High cost products without quality". This review speaks about opinion on two features – positive expression on 'infrastructure' and negative expression on 'cost'. The phrase *'very good'* is associated with the feature infrastructure and *'no satisfaction'*, *'high cost'* is associated with the feature cost.

Most of the web sites rank products or service based on the aggregated user rating given by the user, in spite of the review text given by the user. Moreover, the user also gives random star rating and this may not coincide with the text actually written. It is very much essential to analyze the textual user feedback before ranking or rating any product or service. This paper proposes a sentiment analysis model which involves gathering of user reviews about hospitals of various cities of India from different social media sites. The model analyzes the sentiments of various features expressed in those reviews and gives comparative results of different hospitals. The model also aims to predict the rating of feature of hospital using the semi-automatic text analysis. The proposed model not only helps in decision making for users but also in business intelligence for the promotion of low scored features. It gives the quality indicators of services related to healthcare industry.

This paper is organized as follows: Section 2 discusses the related work, Section 3 details the proposed methodology, Section 4 analyzes the results and Section 5 concludes the paper with the future works explained.

2 Related Work

Social Media has a great impact on all domains and industries. Literature review has been done in three different ways – impact of social media analysis in business or industry, natural language processing techniques for sentiment analysis, feature based sentiment analysis and the use of TF-IDF in sentiment analysis.

2.1 Sentiment analysis and its impact on business

Business like tourism, marketing, healthcare, etc started to utilize technology to reach its target market with the changing technology. Social media becomes a new marketing and communication strategies for all types of industry sectors. Large businesses use social media effectively by analyzing its content. Social media analytics or sentiment analysis on any product or service or business can help in the reputation management of any company or enterprise. Sentiment analysis on the customer feedback will definitely yield significant improvement in the business investment. Voice of customer has been investigated by [Ng, 11] with respect to technical and functional quality of services in the marketing field to improve customer provider relationship. The impact of social media analytics to the economic contribution of industry and thereby to the country was suggested and demonstrated by [Benxiang, 14]. His work also justified that the user generated content in social media web sites are perceived as recommendations from like-minded friends mostly by the younger generations of this century. New conceptual model was developed by [Xin, 14] using Technology Organization Environment theory to asses firm's branding activities using social networking sites. [Olga, 14] studied the evolution of social media and suggested that this platform can be used as the advertizing medium to promote the tourism destinations. [Cheng, 15] analyzed Chinese microblogs on travel sites, travel news and tourists' attitudes to provide an insight of consumers to enable the marketing strategy for Tourism in China. New customer satisfaction model was developed [Padmini, 15] by getting insights from the experience of patients apart from their open ended statements so as to have significant improvement in the healthcare services. The proposed work in this paper analyses the customer feedback written in various social media sites about health care services provided by different hospitals.

2.2 Sentiment analysis using NLP techniques

Machine learning techniques have been applied by [Pang, 02] for extracting subjective portions of document for sentiment classification. [Meena, 07] developed a framework for sentiment analysis for words or phrases in presence of conjuncts using word dependencies and dependency trees. Opinion observer, a linguistic rule based system, was built by [Ding, 07] to detect the polarity of opinions context based and with the use of opinion aggregation function. [Harb, 08] proposed automatic extraction of positive and negative adjectives from blogs and reviews using the association rules for learning the data set and to build the dictionary. Classification accuracy of 71% for positive adjectives and 62% for negative adjectives was obtained by that model. [Damien, 08] developed a sentiment analysis model for film reviews using machine language and NLP techniques. He used improved Naive Bayes classification such as optimal discretization, variable selection, compression based model averaging techniques and used lexicon based approach using different dictionaries. Finally, he proposed that linguistic methods are suitable for smaller data set while machine learning techniques are suitable for larger data set. [Qiu, 08] proposed a method to extract new sentiment words from sentences and assigned polarity for them using double propagation method.

[Jine-Cheon, 10] compared different online genres for moview documents using NLP techniques and identified frequently used positive and negative terms. New method was proposed by [Chawla, 13] for sentiment analysis on tweets having conjunctions, connectives, modals and conditionals using discourse relations for polarity detection of tweets. [Basant, 14] developed a framework based on semantic feature clustering for sentiment analysis using machine learning methods so as to improve the accuracy.

2.3 Feature based Sentiment Analysis

Aspect based sentiment analysis was introduced by [Liu, 12] wherein he explained the structured approach for handling the unstructured data collected through various means. He also gave an in-depth introduction for feature based sentiment classification and comparative opinion mining. [Pontiki, 14] identified the aspect categories for the annotated reviews of laptops and restaurants and evaluated the opinion target expressions of the aspect categories. Multi-perspective classification was proposed by [Duk, 14] to summarize the reviews of a product features. [Chamlertwat, 12] proposed feature based customer feedback analysis on smart phones and summarized the best product feature among various phone brands. [Yang, 10] used context-sensitive information to determine sentiment polarity and opinionated-feature frequency to determine the overall feature scores. [Somprasertsri, 10] developed an ontology based opinion model for mining features of product from customer reviews. [Keefe, 09] used a range of feature selectors like categorical proportional difference (PD), sentiwordnet subjectivity scores (SWNSS), sentiwordnet proportional difference (SWNPD) along with Naïve Bayes and SVM classifiers and confirmed that PD is the best feature selector for the classification of larger data set.

2.4 Sentiment Analysis using TF-IDF

[Vinodhini, 12] surveyed most of feature selection and classification techniques for sentiment analysis for different types of datasets like movie dataset, etc. [Mingyong, 12] proposed improved TF-IDF method for text categorization. They showed 62% and 67% accuracy for 20newsgroup dataset [Jason Rennie, 08] by using Naïve Bayes and SVM classifiers. They used smoothing technique and normalization for TF-IDF weighting method. [Li, 10] used TF-IDF weighing scheme for clustering documents into positive and negative groups. [Xia, 11] used local and global term weight algorithm using TF-IDF values for text categorization. SentiTFIDF model was proposed by [Ghag, 14] which finds the senti-stop-word by using the proportional frequency of the words and by comparing the distribution of words among positive and negative documents in a movie dataset. [Gautam, 13] proposed a new TFIDF-weighting model for sentiment analysis that provides a greater weight to the low frequency words than the frequently occurring words along with the class information. [Martineau, 09] proposed a general purpose technique called DeltaTF-IDF, which gives weight and scores to words before classification. The accuracy of the sentiment analysis is measured by Support Vector Machines.

From the literature it is understood that most of the research focused on sentiment analysis at the document and sentence levels; and very few work has been done on feature based sentiment analysis. They focused only on the techniques themselves, and thus are of limited usefulness for deriving useful information like ranking and selecting of places based on multiple criteria. There is a need to obtain additional information for measuring this. As social media is growing in all directions and makes its foot print in all types of industries, it is very much needed to utilize the technologies for the improvement of industries.

This paper presents a text analytic framework for reviewing user feedback about different hospitals by combining NLP, IR and statistical techniques. The IR technique like TF-IDF can be better utilized in sentiment analysis for feature based sentiment classification and summarization of any product or service. This paper uses TF-IDF and improved TF-IDF methods for aspect level or feature based sentiment analysis. Most of the research used machine learning techniques like naïve bayes, SVM for classification but this paper used lexicon based dictionary approach for classification. The proposed model not only helps in decision making for users but also in business intelligence for the promotion of healthcare industry.

3 Methodology

Sentiment analysis problem is stated as follows: To provide a detailed insight into the customer opinion on different features or aspects of a service or product by analyzing the reviews using NLP techniques for the healthcare domain and to clasiify them into four classes like positive, negative, strong positive and strong negative using feature based sentiment analysis methods. The objective of this paper is to evaluate the strength of the dependencies between the user rating of a review and the different text representations of the review given by the user. The model involves three major steps – (i) data pre-processing (ii) sentiment extraction and bag of words representation and (iii) sentiment classification using regression modeling. Figure 2 shows the proposed model and each step involved in the process are explained in the subsequent sections.

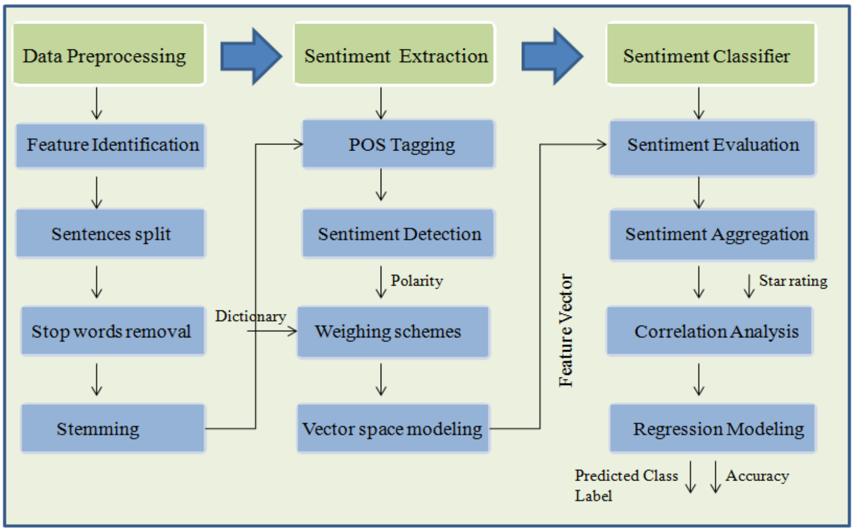

Figure 2: Feature based Sentiment classifier

3.1 Data Pre-processing

Data is collected from various social media sites like twitter using suitable APIs like twitterAPI. Review collection is done based on Keyword search using the name of the place. For example, the keyword ‘Apollo Hospital’ retrieves all tweets from Twitter. Timeframe can also be specified while retrieving the tweets. Reviews are also collected from Google reviews, mouthshut.com, besthospitalAdvisor.com, and the like. It is cleaned and irrelevant review text is removed in order to provide a focus over the useful data. Reviews are extracted for 10 different hospitals under study. These 10 hospitals are identified from the Hospital Ranking websites (like BestHospitalAdvisor) which are popular among the users. Online reviews are collected for the features along with the star rating given by the user. Synonyms of identified features are also used while collecting the review text. Stop words are removed and stemming is done. Sentences are split from the review text for each set of features of products.

3.2 Feature Identification

Features are extracted from the pre-processed data set. In this paper, we have focused on sentiment analysis of features identified in prior. These features are identified based on domain analysis and through user interaction. The document [Robert, 95] also clearly identifies the necessary criteria for the selection of hospital services. The article [Priyanka, 14] identified six different selection criteria like cost, hospital reputation, facility, referrals, effective caring, and medical expertise expected by the patients. Features are also identified by referring web sites like webometrics, BestHospitalAdvisor and so on. These websites analyze the user comments for different features of healthcare services of different hospitals. The most discussed topics are considered as features for this problem, as shown in Table 1.

Features	Synonyms of Features
Cost	location, convenience, expensiveness
Medicare	treatment, therapy, reputation, specialists, quality care, experience
Nursing	attitude, known staff, experience
Infrastructure	facility, cleanliness, equipment, size, technology
Time	quick treatment, less time, modern services

Table 1: Features of Health care services

Synonyms of identified features are used for the separation of reviews. For example, the synonyms of feature ‘Medicare’ include treatment, cure, therapy, etc., the words mostly associated with doctor’s care. Suppose if single review talks about multiple features, the sentences are split based on features and placed in the corresponding location for further analysis. Thus opinions are extracted and grouped for various features of different places.

3.3 Part Of Speech (PoS) tagging

In this step, the sentences in the pre-processed data are tokenized using the PoS tagger. During this process, a Part of Speech (PoS) such as noun, verb, adverb, adjective, conjunctions, negations and the like are assigned to every word in the sentences.

A conjunction is a word that is used to link words, phrases, or clauses. It is often used to express the relationship between two clauses or phrases in a sentence. A double negative always renders a positive effect i.e. not dislike means like. Conjunctions and double negations have the capability to completely invert the polarity or sentiment of the sentence. Analysis is done for conjuncts and negations used in the sentence by observing the position where they occur in the sentence or clause. The rules are applied to the sentence and based on senti-scores of the words in each sentence its sentiment orientation is determined [Askarunisa, 14]. For example, the rules for some of the conjunctions are discussed:

conjunction: though
<conjunction = though >
<rule LC=positive , RC=negative , polarity=positive />
<rule LC=negative, RC=positive , polarity=negative />
</conjunction>

conjunction: but
<conjunction = but >
<rule LC=positive, RC=negative, polarity=negative />
<rule LC=negative, RC=positive, polarity=positive />
</conjunction>

where LC is left clause, RC is right clause, w is word, nw is next word. Consider a sentence with conjunction ‘but’. “The rooms are maintained neatly but the room rent is costly”. In this sentence, the first clause is positive however the clause following the conjunction has a negative effect. So positive score is given for the feature ‘Infrastructure’ and negative score is given for the feature ‘Cost’.

double negation : not <neg adjective>
<double negation>
<rule w=negative , nw=positive , polarity=negative />
<rule w=negative , nw=negative , polarity=positive>
</double negation>

If there are two negation terms occurring in a sentence either consecutively or in different parts of the sentence, the polarity of the sentence is evaluated as positive. Consider another example for double negation. “Nursing care is not bad as I expected”. The negation ‘not’ is followed by the negative adjective ‘bad’ and the polarity is positive. The feature ‘Nursing’ is given +1 score.

3.4 Sentiment Detection

In general, the words like adjective, adverb or verb expresses the sentiment in the sentence. The term frequency matrix is built for each feature under consideration. Based on the category determined by the sentiment dictionary for the sentimental word (positive, negative, strong, weak, pleasure, pain and feel), the weight is assigned for the word i.e. +2 is given for strong positive word, +1 is assigned for positive word, -2 for strong negative word and so on. As an example, the senti-scores of the words ‘very good’ and ‘extremely good’ are assigned +2. The words, ‘very’ is categorized as “overst”, ‘good’ as “positive” and ‘extremely’ as “strong” by General Inquirer. The category ‘Overstated’ indicates realms of speed, frequency, causality, size, etc.

3.5 Vector Space Modelling

Review text is converted into keyword vectors of customer’s opinions. In review text, if feature (or attribute) and sentiment words appear in the same sentence, that sentence is considered to be a customer opinion. In this case, the sentiment score is

counted for the corresponding feature. For example, cosnider a customer writes the following review about a hospital:
*Service was **good**. Doctors were **great, good**, and **expertise.** Treatment was **great**. Quality was visible in their **expertise**. Most of the Doctors have **expertise** in complex operations. But I was **unhappy** on one silly incident which may be avoided in future.*

This review document talks about the feature 'Medicare' of the particular hospital. The adjectives like good, great, expertise and unhappy are extracted. All these sentiment words are related with the feature 'Medicare'. The review documents are represented in BagOfWords (BoW) matrix which is explained in the subsequent section 3.5.1. This section illustrates how this text is transformed into Term-Document matrix using different weighing schemes like Term Frequency (TF), Term Frequency Inverse Document Frequency (TF-IDF) and Term Frequency Inverse Positive Negative Document Frequency (TF-PN-IDF).

3.5.1 Weighing schemes

TF Method:

TF method represents the BoW matrix as the number of occurrences of terms appear in the text. Table 2 shows the six sample review documents written for the feature 'Medicare' for a particular hospital.

Doc No	good	great	expertise	poor	bad	unhappy	user rating	Senti-score (P-N)	Label
1	2	1	3	0	0	1	4	5	pos
2	1	1	1	0	0	0	5	3	pos
3	1	1	2	0	1	0	4	3	pos
4	0	1	0	2	1	1	2	-3	neg
5	0	0	1	1	1	0	1	-1	neg
6	0	0	0	2	2	0	2	-4	neg

Table 2: TF Term-Document Matrix

Senti-score is calculated by number of positive words (P) minus number of negative words (N) i.e. P-N. We can set the class label as 'pos', if the senti-score is positive value and 'neg' if the senti-score is negative value. In other words, if the user rating is 3 and above, the document can be considered as positive document; else it is negative document. In Table 2, the words like good, great, expertise are positive words; words like poor, bad and unhappy are negative words. In this approach, no variation is seen between the documents 2 and 3 through the senti-score, though the user rating is *'5'* and *'4'* respectively. It is better to use 'weights' for the term-frequency based on the type of the term, as explained in Section 3.4.

Table 3 shows the weighted TF matrix for the sample review documents. For example, the dictionary General Inquirer (GI) classifies the term "good" as 'positive', "great" as 'strong positive' and "expertise" as 'strong positive power'. These terms are given the weights +1, +2 and +3 respectively and are multiplied with the frequency of their occurrences as shown in Table 3.

Doc No	**good (+1)**	**great (+2)**	**expertise (+3)**	**poor (-1)**	**bad (-1)**	**unhappy (-2)**	**user rating**	**Senti-score (P-N)**	**Label**
1	2	2	9	0	0	-2	4	11	pos
2	1	2	3	0	0	0	5	6	pos
3	1	2	6	0	-1	0	4	8	pos
4	0	2	0	-2	-1	-2	2	-3	neg
5	0	0	3	-1	-1	0	1	1	pos
6	0	0	0	-2	-2	0	2	-4	neg

Table 3: Weighted TF Term-Document Matrix

A document collection may contain documents of different lengths. It is useful to use normalized weight assignments. Table 3 is normalized and is shown in Table 4. From Table 4, it is evident that the difference in senti-score for the review documents 2 and 3 is clear; also they are in correspondence with the user rating *'5'* and *'4'* respectively.

Doc No	good (+1)	great (+2)	expertise (+3)	poor (-1)	bad (-1)	unhappy (-2)	user rating	Senti-score (P-N)	Label
1	0.21	0.207	0.933	0	0	-0.21	4	1.14	pos
2	0.27	0.535	0.802	0	0	0	5	1.60	pos
3	0.15	0.309	0.926	0	-0.15	0	4	1.23	pos
4	0	0.555	0	-0.55	-0.28	-0.55	2	-0.83	neg
5	0	0	0.905	-0.30	-0.30	0	1	0.30	pos
6	0	0	0	-0.71	-0.71	0	2	-1.41	neg

Table 4: Normalized Weighted TF Term-Document Matrix

TF-IDF Method:

Though the user rating is *'4'* for two documents 1 and 3, the presence of weak negative word 'bad' and strong negative term 'unhappy' is visible through the senti-score values *1.14* and *1.23* respectively, as shown in Table 4. But Table 4 shows the senti-score of document 5 as positive and is labeled as 'pos'; it is not aligned with its user rating *'1'* and it is supposed to be 'neg'. TF-IDF method may be quite helpful to overcome this issue – to identify the presence of negative terms in the document, in spite of presence of positive terms. In this method, TF-IDF values are used to represent BoW matrix. TF-IDF adds weighting to a term based on its inverse document frequency. It means that if the term appears in more documents, lesser the importance is and the weighting will be less. It can be depicted as:

$$score\ (tfidf) = f_{ij} * \log \frac{N}{n_j} \qquad (1)$$

$$Normscore\ (tfidf) = \frac{f_{ij} * \log\left(\frac{N}{n_j}\right)}{\sqrt{\sum_{j=1}^{M}\left[f_{ij} * \log\left(\frac{N}{n_j}\right)\right]^2}} \quad (2)$$

where f_{ij} represents the term frequency of term 'j' in review document 'i', N represents the number of reviews and n_j represents the number of documents that the term j occurs.

Table 5 shows the normalized weighted TF-IDF Term Document matrix using the equations (1) and (2). Still, the document 5 has positive senti-score value but its value is reduced to 0.0808. This is due to the fact that the frequency of occurrence of terms 'expertise' and 'bad' are same in all 6 document collection.

Doc No	good (+1)	great (+2)	expertise (+3)	poor (-1)	bad (-1)	unhappy (-2)	user rating	Senti-score (P-N)	Label
1	0.3045	0.1781	0.8016	0	0	-0.4827	4	0.8016	pos
2	0.4284	0.5012	0.7518	0	0	0	5	1.6815	pos
3	0.2579	0.3018	0.9053	0	-0.1509	0	4	1.3142	pos
4	0	0.2947	0	-0.5038	-0.1474	-0.7985	2	-1.1549	neg
5	0	0	0.8345	-0.4755	-0.2782	0	1	0.0808	pos
6	0	0	0	-0.8632	-0.5049	0	2	-1.3680	neg

Table 5: Normalized Weighted TF-IDF Term-Document Matrix

TF-PN-IDF method:

The disadvantage of TF-IDF can be overcome by using improved TF-IDF method 'TF-PN-IDF'. In this method, the frequency of occurrence of terms in the corresponding class (positive or negative) is also considered. In other words, TF-P-IDF (Term Frequency Inverse Positive Document Frequency) and TF-N-IDF (Term Frequency Inverse Negative Document Frequency) reflects the importance of the term in the positive and negative documents respectively. TF-PN-IDF is composed of two parts:

$$score\ (tfpidf) = f_{ki} * \frac{P_i}{S_p} \log \frac{N}{n_i}, (document_k \in P)$$
$$score\ (tfnidf) = f_{ki} * \frac{N_i}{S_n} \log \frac{N}{n_i}, (document_k \in N) \quad (3)$$

where f_{ki} is the frequency of term i in document k, N is the number of documents in the collection, n_i is the number of documents where term i occurs in the collection, P_i is the number of positive documents where term i occurs, N_i is the number of negative documents where term i occurs, S_p and S_n are the number of positive and negative documents in the collection, respectively. A vector length normalization of TF-PN-IDF is defined by:

$$score_{ki} = \frac{f_{ki} * \frac{P_i}{S_p} * \log\left(\frac{N}{n_i}\right)}{\sqrt{\sum_{r=1}^{M}\left[f_{kr} * \frac{P_r}{S_p} * \log\left(\frac{N}{n_r}\right)\right]^2}}, document_k \in P$$
$$score_{ki} = \frac{f_{ki} * \frac{N_i}{S_n} * \log\left(\frac{N}{n_i}\right)}{\sqrt{\sum_{r=1}^{M}\left[f_{kr} * \frac{N_r}{S_n} * \log\left(\frac{N}{n_r}\right)\right]^2}}, document_k \in N \quad (4)$$

where M is the number of all terms in the review collection. Table 6 shows the normalized weighted TF-PN-IDF values of each term for each document. Using the approach TF-PN-IDF, the document 5 takes the label ‘neg’ as its senti-score is negative.

Doc No	good (+1)	great (+2)	expertise (+3)	poor (-1)	bad (-1)	unhappy (-2)	user rating	Senti-score (P-N)	Label
1	0.342	0.201	0.902	0	0	-0.181	4	1.262	pos
2	0.429	0.501	0.752	0	0	0	5	1.682	pos
3	0.261	0.305	0.915	0	-0.051	0	4	1.429	pos
4	0	0.165	0	-0.844	-0.247	-0.446	2	-1.373	neg
5	0	0	0.451	-0.771	-0.451	0	1	-0.771	neg
6	0	0	0	-0.863	-0.505	0	2	-1.369	neg

Table 6: Normalized Weighted TF-PN-IDF Term-Document Matrix

3.6 Sentiment Aggregation

Words are grouped into positive and negative categories. Senti-score of each review text for each feature is determined by

Senti-score = P - N (5)

where P represents the score of all positive words and N represents the score of all negative words. The aggregated value of P-N of all features of product or service gives the overall senti-score of the product or service.

3.7 Correlation Analysis

Correlation analysis is done between the senti-score and the star rating. Strong correlation says that the user consciously marks the overall star rating. Weak correlation may state that the user randomly marks the star rating. Correlation analysis between these two methods helps in identifying the suitable BoW representation method for further predictive modelling.

3.8 Sentiment Classifier

This module classifies the sentiment text of each feature of each review into four different classes like positive, negative, strong positive and strong negative using the linear regression modelling technique. Let $\mathbf{X_{ij}}$ be the independent variable, the senti-score of j^{th} feature of i^{th} product or service and Y be the dependant variable, the sentiment category of the sentence. The sentiment category is determined by the regression model:

$$Y = \mu + ß_i X_{ij} \tag{6}$$

where µ is the error term and ß is the estimate value, which gives the sentiment class of the text like positive, negative, strong positive or strong negative. This is further explained in detail in Section 4.3.

4 Results and Discussion

The proposed model has been implemented using social media related APIs, NLP tools and IR techniques. Initially, the user generated data for remarkable hospitals are collected using twitterAPI, Google reviews, etc. Reviews are categorized based on the places. The reviews are collected for different features of various hospitals and are shown in Table 7 and 8. The rating of a review may be 1, 2, 3, 4 or 5. Some websites provide feature based rating for the product or service. Approximately, 25% reviews have user rating 4 or 5, 45% reviews have rating 3, 18% reviews have rating 2 and 12% reviews have rating 1. Labeling of these reviews is not done manually. For the experiment, reviews with the user rating 3, 4 and 5 are considered as 'positive' label and reviews with the user rating 1 and 2 are considered as 'negative' label.

Data Source	**Number of Reviews**	
	Positive	Negative
Twitter	1200	525
Mouthshut.com	425	110
BestHospitalAdvisor.com	200	85
Google Reviews	580	320

Table 7: Dataset collection of online reviews

Data collection shows that people use social media rigorously during the period 2014 and 2015. Table 8 shows the number of reviews collected for different hospitals and are referred by H1, H2, and so on.

Features/ Places	**H1**	**H2**	**H3**	**H4**	**H5**	**H6**	**H7**	**H8**	**H9**	**H10**
Cost	114	103	81	51	23	99	54	21	38	79
Medicare	165	75	142	75	45	33	71	47	61	34
Nursing	174	41	74	25	91	64	53	77	92	85
Infrastructure	55	47	69	35	19	73	51	81	46	78
Time	88	145	33	41	74	39	84	71	44	85
Number of reviews	596	411	399	227	252	308	313	297	281	361

Table 8: Number of online reviews for different places

Feature extraction phase includes the identification of most talked topics and their attributes. For example, if a tweet states *"Very quick procedures. I had to get many tests done, but they took far less time than I had expected"*, then this step extracts feature 'Time' by understanding the synonymous words quick, less time, etc. Another example states *"The hospital wards are very clean and neatly maintained. Good care on patients. Spent only few days"*, then this step extracts features like Infrastructure, Medicare, Time, etc. Table 9 shows the weight calculation of a simple positive sentence *"Doctors are excellent"*.

Words	**POS Tagging**	**Meaning**	**GI Category**	**Weight**
Doctors	/NN	Noun	-	0
Are	/VBZ	Verb	-	0
Excellent	/JJ	Adjective	Strong and Positive	2

Table 9: POS tagging of a sentence

In another example, *"Doctors at this hospital do a much better job of diagnosing illness and they are excellent"*, there are positive words like "much", "better" and "excellent". Using General Inquirer (GI) dictionary, the senti-score for each word is assigned. If the word is positive, it is assigned weight +1; if the word is strong and positive, it is assigned +2. Hence, overall score for this sentence is 4 for the feature 'Medicare'.

4.1 Experimental Results

During the training phase, the rules have been identified for sentiment orientation considering the presence of conjuncts, to improve the accuracy of decision on tweets and opinions. Different reviews are considered and scores are given to each feature by sentiment evaluation and aggregation. During the evaluation phase, the TF value is determined for each sentiment word present in the text for each feature as shown in Table 10 for the hospital H1 [as explained in Table 4]. Using the General Inquirer dictionary, the sentiment words are given weights if they are strong or weak, as described in Sections 3.4 and 3.5 and the senti-score is calculated using Equation (5).

DocNo	**Cost**	**Medicare**	**Nursing**	**Infrastructure**	**Time**	**Overall score**	**UserRating**
1	0.13	1.14	0.38	0.18	0.12	1.95	4.00
2	0.00	1.60	0.12	0.32	0.23	2.27	5.00
3	1.45	1.23	0.12	0.00	-1.34	1.46	4.00
4	0.00	-0.83	0.56	-0.23	0.65	**0.15**	2.00
5	0.00	0.30	-0.48	0.00	0.98	**0.80**	1.00
6	0.00	-1.41	-0.13	0.30	0.00	-1.24	2.00
Total	1.58	2.03	**0.57**	**0.57**	0.64		

Table 10: Term-Frequency for features and reviews

TF-IDF value for each review document is calculated using the equations (1) and (2). Sentiment orientation is done for each feature for hospital H1 as shown in Table 11 [as explained in Table 5].

Doc No	Cost	Medicare	Nursing	Infrastructure	Time	Overall score	User Rating
1	0.29	0.80	0.83	0.93	0.08	2.93	4.00
2	0.00	1.68	0.24	0.15	0.58	2.65	5.00
3	0.98	1.31	0.34	0.00	-0.14	2.49	4.00
4	0.00	-1.15	0.75	-0.84	0.85	-0.39	2.00
5	0.00	0.08	-0.25	0.00	0.73	**0.56**	1.00
6	0.00	-1.37	-0.26	0.16	0.00	-1.47	2.00
Total	1.27	1.35	1.65	0.40	2.10		

Table 11: TF-IDF for features and reviews

TF-PN-IDF value for each review document is calculated using the equations (3) and (4). Sentiment orientation is done for each feature for hospital H1 as shown in Table 12 [as explained in Table 6].

Doc No	Cost	Medicare	Nursing	Infrastructure	Time	Overall score	User Rating
1	0.19	1.26	0.43	0.36	0.08	2.32	4.00
2	0.00	1.68	0.24	0.15	0.58	2.65	5.00
3	0.78	1.43	0.34	0.00	-0.14	2.41	4.00
4	0.00	-1.37	0.75	-0.44	0.85	**-0.21**	2.00
5	0.00	-0.77	-0.72	0.00	0.43	**-1.06**	1.00
6	0.00	-1.37	-0.14	0.53	0.00	**-0.98**	2.00
Total	0.97	0.86	0.90	0.60	1.80		

Table12: TF-PN-IDF for features and reviews

Table 10 shows the aggregated senti-score for the features ‘Nursing’ and ‘Infrastructure’ as *0.57*. This value does not give any good judgment between these features. When TF-IDF is used for calculating the senti-scores, as shown in Table 11, it is evident that for the hospital H1, the feature ‘Infrastructure’ *(20.52)* is better than ‘Nursing’ *(11.56)*. Table 10 shows the overall senti-score of the hospital H1 by the review text R4 and R5 as positive. They are supposed to be negative w.r.t. user rating. But Table 11 shows good differences in the overall score of review text R4 *(-0.39)*, still the overall score of R5 shows positive. This issue is overcome by the weighing scheme TF-PN-IDF method. Table 12 shows the overall score of all six documents, which are aligned with the user rating.

Further, the expressiveness or the feelings of users can be really measured by the TF-IDF methods not by simple TF method. The use of particular word in the entire document collection is clearly brought by IDF component. For example, the word ‘good’ is often used in most of the documents. But the words like ‘extraordinary’, ‘awesome’, etc., are used in few documents. These words should be given more weight subsequently. So, TF-IDF and TF-PN-IDF methods enable better estimating of senti-scores for features of product or service.

4.2 Statistical Test

The relationship between senti-score and the star rating can be established using correlation analysis of statistical tests. Pearson Correlation is widely used to measure the relationship degree between the two variables. Correlation is significant when *p* value is less than hypothetical error level (0.01 or 0.05). Null hypothesis is that there is no relation between the senti-score and the star rating of a place. Table 13 shows the correlation co-efficient '*r*' and the '*p*' value. It is evident from the '*r*' value that TF-PN-IDF method has stronger correlation between the variables than other two weighing methods. Also the '*p*' value shows that the two variables are statistically significant.

Methods	**Correlation Co-efficient (*r*)**	***p* value (at 0.05)**
Normalized Weighted TF	0.76	0.04
Normalized Weighted TF-IDF	0.83	0.009
Normalized Weighted TF-PN-IDF	0.92	0.005

Table 13: Correlation analysis

Scatter plot drawn between the senti-score and the star rating for sample data *(around 100 reviews)* is shown in Figure 3. Normalized weighted TF-PN-IDF weighing scheme is used for calculating the senti-score.

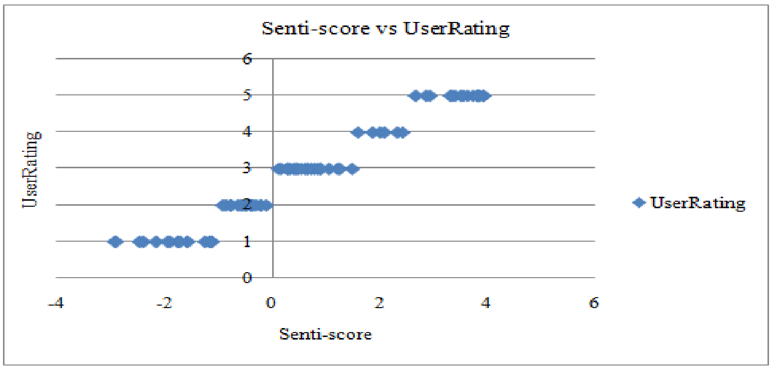

Figure 3: Senti-score vs StarRating

From Figure 3, it is evident that the senti-score values can be easily differentiated and classification can be performed as follows:

if, *senti-score < -1, the class label is Strong Negative (User rating 1)*
if, *-1 < senti-score < 0, the class label is Negative (User rating 2)*
if, *0 < senti-score < 2, the class label is Positive (User rating 3)*
if, *senti-score > 2, the class label is Strong positive (User rating 4 and 5)*

But this approach would be better, if the data size is very small. Also, rating *'4'* and *'5'* are not distinguished.

4.3 Linear Regression Model

Regression modeling is the suitable technique for predicting the class label, if the data set size is larger. Linear regression is run between the user rating (dependent variable) and the senti-score (independent variable), using the equation (6). Regression is run using three different weighing schemes for all the review documents that are under consideration. The weighing scheme TF-PN-IDF along with the regression analysis on the two variables gives the linear regression model as:

$$\text{UserRating} = \mu + ß_i * \text{senti-score}$$

where µ takes the value 2.87 and $ß_i$ takes the value 0.031.

Prediction of test review data set is done using this relationship equation and the accuracy is calculated as shown in Table 14. Method 1 uses normalized weighted TF, Method 2 uses normalized weighted TF-IDF and Method 3 uses normalized weighted TF-PN-IDF as the weighing schemes. Significant improvement is seen in the result when TF-PN-IDF is used as the weighing scheme.

Data Set		**Accuracy (in %)**		
Train	Test	Method 1 (TF)	Method 2 (TF-IDF)	Method 3 (TF-PN-IDF)
3000	350	82	88	93
2000	1350	73	81	85
1500	1850	65	74	82
1000	2350	52	67	78

Table 14: Accuracy of Predictive Model

Table 15 shows the comparative results of classification for different variations of TF techniques. The weighing scheme, 'Normalized weighted TF-PN-IDF' shows the better result for all classification methods like Naïve Bayes, SVM and regression. The TF-IDF method differentiates the different classes but TF-PN-IDF method has the ability to represent that class itself. The term will be given more importance if that term appears more number of times in the document of that class. This results in better accuracy of different classification methods.

Methods	**Weighting scheme**	**Accuracy**		
		Naive Bayes	**SVM**	**Regression**
[Mingyong, 12] for 20Newsgroup dataset	Normalized TF-IDF	62%	67%	
	Improved TF-IDF (using in class frequency)	77%	78%	
Proposed Method	Normalized Weighted TF-IDF	67%	74%	78%
	Normalized Weighted TF-PN-IDF	75%	81%	85%

Table 15: Comparison of results

The senti-score of each feature of each place is aggregated and is shown in Figure 4. The place H10 ranks first for nearly three features like Cost, Medicare and Infrastructure. It still needs improvement for the features like Nursing and Time. Thus ranking of features based on the senti-score improve decision making for both business and other users.

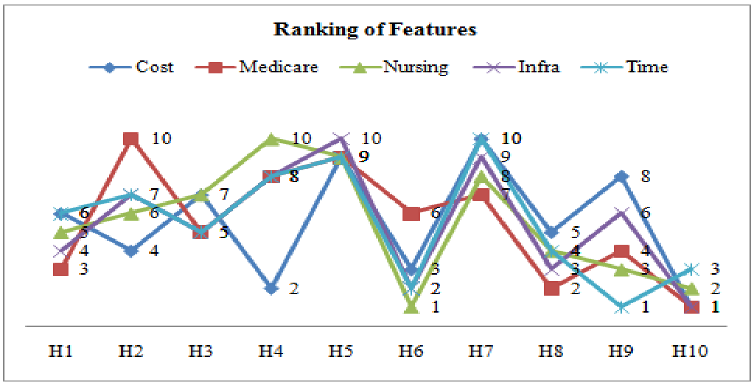

Figure 4: Ranking of Features

The relationship between the ranks given by the proposed method and the popular web site is shown in Table 16. The web site might have used star ratings of the user, but the proposed method used the review text for determining the senti-scores of features.

Hospitals	Website Ranking	Ranking by proposed method
H1	4	5
H2	9	6
H3	5	7
H4	10	10
H5	7	9
H6	2	2
H7	6	8
H8	8	4
H9	3	3
H10	1	1

Table 16: Ranking of Hospitals

5 Conclusion

The reports from social media analytics of service providers like healthcare proved that the reputation and quality improvement go hand-in-hand. So it becomes necessary to manage online reputation to maintain the trust with clients. This study proposed the sentiment analysis framework for comparing features of different hospitals based on the online reviews generated by the users. The contribution and potential utility of this methodology is twofold. First, customer surveys on products or services are used to measure the customer satisfaction level traditionally. More time and manual effort is spent in analyzing these surveys. To solve this problem, the proposed approach measures the customers' feelings through sentiment analysis methodology. The proposed approach is more effective as it uses the textual representation of customer views directly. The analysis on user generated data provides a higher degree of accuracy about what exactly the user feels about a particular place. Feature based analysis recommends the feature of a particular place either it is positive or negative. In this paper, we have presented lexicon based approach for sentiment analysis of online reviews collected from various social media web sites. Specifically, we analyzed different features of hospitals using the reviews generated by the users. Second, this study used TF-IDF, the IR technique for the bag of words representation of user reviews and linear regression is used as the classifier model. The accuracy of sentiment classifier is improved when improved TF-IDF method i.e. TF-PN-IDF is used for representing the review text. Also the statistical correlation tests prove that TF-PN-IDF is the better weighing scheme for representing the review text.

The focus of this study was not limited to the means of measuring customer satisfaction; rather, the proposed approach provides guidelines for identification of areas for improvement in specific aspects of service operation. Feature based sentiment analysis helps to identify the feature which has strong negative sentiment among users. It can be addressed at the right time so as to implement new business intelligence solutions and decision making for emergency readiness, overall improvement and so on. Most of the web sites rate products or services based on the user rating given to them. But the proposed method considered the user generated content also. So, this type of feature based sentiment analysis or user generated content analysis facilitates decision-making process for different types of patients and healthcare providers. Though the proposed method is able to identify <feature, opinion> pairs, identifying sarcasm type of statements used in the review text is challenging and complex. This can be done by building the domain based context aware sentiment dictionary.

The future work includes the complete automation of sentiment analysis for any domain or industry. Latent Dirichlet Allocation can be used for the automatic extraction of features from online reviews. Knowledge discovery, domain based dictionary and query processing can be enhanced by the use of knowledge management techniques like ontology and semantic web technologies. All these work enables the full automation of sentiment analysis framework for any product, service or industry.

References

[Askarunisa, 14] Askarunisa, A., Abirami A.M.: “Resolving conjuncts in sentiment Analysis for online customer reviews”, Proceedings in 2nd International Conference on Business Analytics and Intelligence (2014)

[Basant, 14] Basant, A., Namita, M.: “Semantic Feature Clustering for Sentiment Analysis of English Reviews”, IETE Journal of Research, 60, 6 (2014), 414-422.

[Benxiang, 14] Benxiang, Z., Gerritsen, R.: “What do we know about social media in tourism? A review”, Tourism Management Perspectives, 10 (2014), 27-36.

[Best Hostpital Advisor website, 15] Retrieved from www.besthostpitaladvisor.com (July 2015)

[Chawla, 13] Chawla, K., Ramteke, A., Bhattacharyya, P.: "IITB-Sentiment-Analysts: Participation in Sentiment Analysis in Twitter SemEval 2013 Task", Seventh International Workshop on Semantic Evaluation (2013), 495-500.

[Chamlertwat, 12] Chamlertwat, W., Bhattarakosol, P., Rungkasiri, T., Haruechaiyasak, C.: "Discovering Consumer Insight from Twitter via Sentiment Analysis", Journal of Universal Computer Science, 18, 8 (2012), 973-992.

[Cheng, 15] Cheng, M., Edwards, D.: "Social Media in Tourism - a visual analytic approach", Current Issues in Tourism, 18, 11 (2015), 1080-1087.

[Damien, 08] Damien, P., Bothorel, C., Guimier De Neef, E., Boull´e, M.: "Automating Opinion analysis in Film reviews: the case of statistic versus Linguistic approach", Language Resources and Evaluation Conference, Morocco (2008), 94-101.

[Ding, 07] Ding, X., Liu, B.: "The Utility of Linguistic Rules in Opinion Mining", Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development on Informations Retrieval, SIGIR'07, Amsterdam (2007)

[Duk, 14] Duk, K.H., Kavita, G., Parikshit, S., Xiang, Z.C.: "Comprehensive Review of Opinion Summarization", International Journal of Computer Engineering and Applications (2014)

[Gautam, 13] Gautam, J., Kumar, E.: "An Integrated and Improved Approach to Terms Weighting in Text Classification", International Journal of Computer Science Issues, 10, 1 (2013)

[Ghag, 14] Ghag, K., Shah, K.: "SentiTFIDF – Sentiment Classification using Relative Term Frequency Inverse Document Frequency", International Journal of Advanced Computer Science and Applications, 5, 2 (2014)

[Harb, 08] Harb, A., Plantié, M., Dray, G., Roche, M., Trousset, F., Poncelet, P.: "Web Opinion Mining: How to extract opinions from blogs?", CSTST '08 International Conference on Soft Computing as Transdisciplinary Science and Technology, (2008), 211-217.

[JagadeeshKumar, 14] JagadeeshKumar, M.: "Expanding the boundaries of your research using social media: stand-up and be counted", IETE Technical Review, 31, 4 (2014), 255-257.

[Jine-Cheon, 10] Jine-Cheon, N., Thet, T.T., Khoo, C.G.: "Comparing sentiment expression in movie reviews from four online genres", Online Information Review, 34, 2 (2010), 317-338.

[Keefe, 09] Keefe, T., Koprinska, I.: "Feature Selection and Weighting Methods in Sentiment Analysis", Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia (2009)

[Li, 10] Li, G., Liu, F.: "Application of clustering method on sentiment analysis", Journal of Information Science, 38, 2 (2010), 127-139.

[Liu, 12] Liu, B.: "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers (2012)

[Martineau, 09] Martineau, J., Finin, T.: "Delta TFIDF - an Improved Feature Space for Sentiment Analysis", Third AAAI International Conference on Weblogs and Social Media, San Jose CA (2009)

[Meena, 07] Meena, A., Prabhakar, T.V.: "Sentence Level Sentiment Analysis in the Presence of Conjuncts Using Linguistic Analysis", 29th European Conference on IR Research ECIR 2007, LNCS 4425 (2007), 573–580.

[Mingyong, 12] Mingyong, L., Yang, J.: "An improvement of TFIDF weighting in text categorization", International Proceedings of Computer Science and Information Technology, 47, 9 (2012)

[Mouthshut website, 15] Retrieved from www.mouthshut.com/product/categories/hospitals (July 2015)

[Newsgroups dataset, 08] http://qwone.com/~jason/20Newsgroups/

[Ng, 11] Ng, S., David, M.E., Dagger, T.S.: "Generating positive word-of-mouth in the service experience", Managing Service Quality, 21, 2 (2011), 133-151.

[Olga, 14] Olga, L.P, Raj, R.: "Evolution of social media and consumer behaviour changes in tourism destination promotion", International Journal of Business and Globalisation, 12, 3 (2014), 358–368.

[Padmini, 15] Padmini, V., Tannirua, M.: "Seeking intelligence from Patient experience using text mining: analysis of emergency data", Information Systems Management, 32, 3 (2015), 220-228.

[Pang, 02] Pang, B., Lee, L.: "Thumps up? Sentiment Classification using Machine Learning techniques", Proceedings of Empirical Methods in Natural Language Processing (2002), 79-86.

[Pontiki, 14] Pontiki, M., Galanis, D., Papageorgiou, H., Manandha, S., Androutsopoulos, I.: "SemEval-2014 Task 4: Aspect Based Sentiment Analysis", Proceedings of the 8th International Workshop on Semantic Evaluation, Dublin, Ireland (2014), 27–35.

[Priyanka, 14] Priyanka, A., Janssens, B.: "Power to the patient: A new growth paradigm for Indian providers", BCG perspectives (2014)

[Qiu, 08] Qiu, G., Liu, B., Bu, J., Chen, C.: "Expanding Domain sentiment lexicon through double propagation", Computational Linguistics, 37, 1 (2008), 9-27.

[Robert, 95] Robert, J.: "An examination of consumer criteria for choosing hospital services" (1995)

[Somprasertsri, 10] Somprasertsri, G., Lalitrojwong, P.: "Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization", Journal of Universal Computer Science, 16, 6 (2010), 938-955.

[Twitter website, 15] Retrieved from https://dev.twitter.com/docs/using-search (July 2015)

[Vinodhini, 12] Vinodhini, G., Chandrasekaran, R.M.: "Sentiment Analysis and Opinion Mining: A survey", International Journal of Advanced Research in Computer Science and Software Engineering, 2, 6 (2012)

[Xia, 11] Xia, T., Chai, Y.: "An improvement to TF-IDF term distribution based term weight algorithm", Journal of Software, 6, 3 (2011), 413-420.

[Xin, 14] Xin, J.Y., Ramayah, T., Soto-Acosta, P., Popa, S., Ping, T.A.: "Analyzing the Use of Web 2.0 for Brand Awareness and Competitive Advantage - an Empirical Study in the Malaysian Hospitability Industry", Information Systems Management, 31, 2 (2014), 96-103.

[Yang, 10] Yang, J., Kim, H., Le, S.: "Feature based product review summarization utilizing user score", Journal of Information Science and Engineering, 26 (2010), 1973-1990.