ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(19), DOI: 10.17485/ijst/2016/v9i19/93876, May 2016* ISSN (Online) : 0974-5645

Sentiment Mining from Online Patient Experience using Latent Dirichlet Allocation Method

P. Padmavathy and A. Anny Leema

Department of Computer Applications, B. S. Abdur Rahman University, Chennai - 600048, Tamil Nadu, India;
padmavathy28@bsauniv.ac.in, annyleema@bsauniv.ac.in

Abstract

Background/Objectives: The paper is focusing on the problem of mining sentiments in the health care text. The attempt here is to apply sentiment analysis technique to extract the feelings of patients with various emotion labels like happiness, sadness and surprise about the healthcare. **Methods/Statistical Analysis:** In this paper the connectivity between social emotions and affective terms are predicted from the patient experience automatically using a joint emotion-topic model by augmenting Latent Dirichlet Allocation (LDA) along with a layer for emotion modeling. The following six modules like Preprocessing, Topic Generation, Polarity Classification, Sentiment Classification, Sentiment Analysis and Aspect Ranking are identified in our system. The set of latent topics is generated from emotions initially. From each of the latent topic affective terms are generated. Finally K-means clustering is applied to detect the emotion. Using aspect ranking technique the weightage of the document is calculated. **Findings:** An intricate description about sentiments reflected in the reviews of patient experience is not provided by many of the sentiment prediction approaches. Experimental results proved that the meaningful latent topics for each emotion are successfully identified by the proposed model. The identified emotions are useful to categorize the document and assist the online users to select required healthcare based on their emotional preferences. **Application/Improvements:** The machine learning process is able to make a careful determination of patient opinion about the various administration aspects of a hospital based on the prediction accuracy that have been achieved. Various machine learning predictions are correlated with results of more conventional surveys. It will be interesting to generate more efficient algorithms based on topic models in several other opinion mining systems and for large-scale data sets.

Keywords: K-means, LDA, Patient Experience, Sentiment Analysis, Topic generation

1. Introduction

Opinion Mining or Sentiment analysis comprises the building of a system that explores the user opinions that are made in blog posts, comments, reviews or tweets about a product or a topic. The mental state of a user related to a topic is determined. Opinion Mining is a subfield of data mining. It is used to judge the sentiments of the human kind on the web through reviews.

The Web is a huge depository of structured and unstructured data. The study of this data to select the elemental user opinion and sentiment is a provoking task. An opinion can be described based on the four components such as Topic, Holder, Claim and Sentiment[1].

Getting the picture about the patients experience in health care is a thorough process of providing care andit is an indicator of health care quality. Nowadays patients have begun to report their health care experience on the Internet through blogs, social networks and wikis and on health care rating websites. Natural language processing of massive datasets, including Sentiment analysis and opinion mining has been pivotal to understand consumer behavior. It has been proved patient comments about specific doctors could have positive or negative sentiment. It is also suggested that capturing of sentiments should be related to common methods of evaluating patient experience. The Information Strategy for the National Health Service (NHS) in England states that sentiment analysis of data is the initial source of valuable information for patients in assisting to choose hospitals.

* *Author for correspondence*

This NHS website acknowledge the patients to convey about their treatment experiences at various hospitals through online reviews in the form of text descriptions. An opportunity is provided to estimate the accuracy of sentiment analysis techniques against the patients' own quantitative ratings. In sentiment deduction, an opinion mining featuremarks an object or an attribute of an entity on which users voice their opinions. In this paper, we propose a best technique that identifies such features from unstructured opinions[2]. Many approaches are illustrated to extract opinion features in topic generation. Topic modeling groups terms of the same topic into one group. Topic modeling methods are seen as clustering algorithms that group terms into homogeneous topics. Rather than considering each document as "a-bag-of- words" as in most of the models that deal with text documents, topic modeling presumes that a document is "a-bag-of-topics". Topic modeling groups every term inside the document into a proper topic. Topic modeling strategies will perform mining on aspects that are characteristics of the precise opinions mentioned in reviews. To boost the performance of resultant tasks such as social emotion prediction, it is required to accurately illustrate the connections between words and opinions. An additional layer of opinion mining is done using Latent Dirichlet Association to enhance efficiency of the processes. LDA is a generative model that allows documents to be explained by unidentified topic[3,4]. Opinion mining consists of extraction of opinions and polarity from a text expressed by its author. Lexical resources such as polarized lexicons are required for this task. Existing approach is based on the Sentimental Analysis of common discussion. It uses the technique of the sentiment-term model. The sentiment term model treats all the terms individually and does not support for the health care industry. Our technique is briefed as follows: The review content is pre-processed. Stop word removal and stemming are the processes involved in pre-processing. In case of stop word removal the words like a, an, are removed. Identification of the root word is done by applying the Porter stemming algorithm. The POS tagging is performed. The words are tagged as noun, verb and adjective. The topics are identified using Latent Dirichlet Allocation (LDA). The sentiments present are identified from the topics. Using the tagged text, the part of speech is identified from the available sentiments. The clustering technique is applied on the sentiment words.

2. Problem Definition and Analysis

The basic elements of an opinion is the opinion holder that defines the person or organization holding the opinion on selected product; Product is the object on which the opinion is expressed and ultimately opinion is the views, emotions or analysis that has been done on an object by the opinion holder.

2.1 Opinion Mining Tasks

The opinion mining tasks are involved at three levels. They are the document level, sentence level and characteristic level. The classification of reviews is performed based on sentiments[5]. It is determined that every opinion document focuses on one product and contains the opinions from a single opinion holder. The sentence level opinion mining is to identify the subjective/opinionated sentences. It is believed that a sentence will hold only one opinion. Phrases and clauses are also considered. Identifying and selecting the object features that have been criticized by an opinion holder is the primary focus at the feature level. It is necessary to see whether the opinions on the features are positive, negative or neutral. Finally, opinion features are grouped. This is helpful for identifying opinion holders. Sentiment classification is predominantly used to classify the documents based on the thorough sentiments expressed by the opinion holders. Categorization is done as positive, negative and neutral. As per this model an object A itself is a feature set. Opinion mining will pin down the sentiment classification performed on Object A in each document. The topic words are important in topic-based modeling. In case of sentiment classification, sentiment words play a major role, e.g., good, wonderful, amazing, horrible, awful etc.

The Clustered Emotion-Topic model identifies the topics and emotion and may be productively applied to study the social emotions corresponding to the reviews on real world event.

2.2 Problem Statement

The major problem is the discovery of affiliation between social emotions and online documents. The categorization needs to be domain specific as it would diminish the effectiveness of mined results to users.

The issues relating to the problem are as follows:

- Reviews are mostly demarcated as positive and negative. An intricate description about sentiments reflected in the reviews is not provided by many of the sentiment prediction approaches.
- The emotion prediction models the documents under the "bag-of-words" assumption such that the relativity between the words is not taken into consideration.
- Classifiers which work well in one sphere will often arrive to yield anticipated results when applied in another sphere.
- Various models like Support Vector Machine (SVM) and Latent Semantic Analysis (LSA) are supervised.
- The review quality which attains the correctness of the prediction is not processed.

3. Existing Problems

In prior system, the discovery and mining of connections between various social emotions and online documents has been a major drawback. The disadvantage is that the existing system provides the sentimental analysis for the normal discussion and does not support the health care industry. Classification tends to be domain specific and it limits the efficiency of mined results to users. The terms are treated on an individual basis by the opinion model and cannot ascertain the related information within the document. Like all opinions, sentiment is inherently subjective from person to person. The existing system is based on the Sentimental Analysis of common discussion. It uses the technique of the sentiment-term model. The sentiment term model treats all the terms individually. There are several algorithms that are used to implement the techniques related to sentiments. Pattern boot strapping algorithm is used in the[1,2,5] existing system.

4. Proposed System

The objective here is to precisely model the connectivity between words and opinions to upgrade the performance of the social emotion prediction. To elevate the efficiency of the processes LDA, Clustered Emotion-Topic model is used in this work. Latent Dirichlet Allocation is based upon the belief that documents are blend of topics where each topic is a probability distribution over the set of words[6]. In this approach there are three stratified layers in which the topics are associated with documents and the words with topics.

As illustrated in Figure 1, the text is pre-processed. The processing consists of stop word removal and stemming. In case of stop word removal, the words like a, an and are removed. The root word identification is done by implementing Porter stemming algorithm. Then POS tagging is performed and the words are tagged as noun, verb or adjective. Next the topics are detected using the LDA. The sentiments are identified from the topics. The tagged text is also used to pinpoint the part of speech where the sentiments are present. The clustering technique is exercised to the sentiment words and the contextual informationis anticipated.

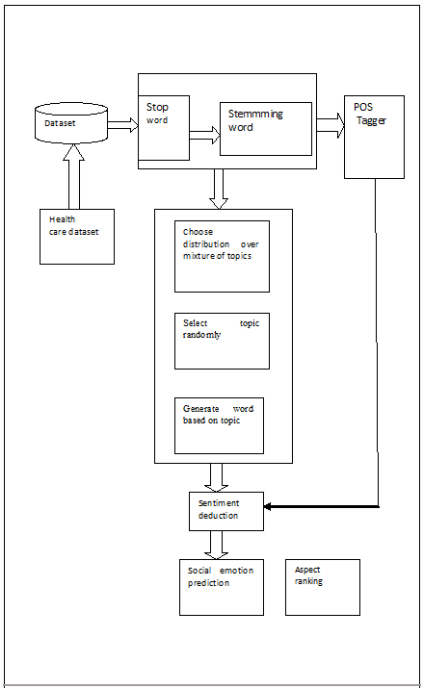

Figure 1. Architecture diagram.

5. Projected Methodology

Several approaches are available in opinion mining. LDA and Clustered Emotion-Topic model are utilized for this paper. LDA model is predicted on the idea that documents are mixture of topics in which each topic is a probability distribution over words.

Three ordered layers are available in LDA, where topics are related with documents and words are associated with topics.

One more layer of emotion modeling is introduced

into the LDA by the Clustered Emotion-Topic Model where the topics are detected and the sentiments are identified. Finally K-means clustering is applied to detect the emotion.

The six modules identified for our system are as follows:

5.1 Preprocessing Module

This module consists of stop word removal and stemming. The common method for removal of words that occur frequently and has no meaning is named as Stop word removal. Stop words do not carry any information. They are language specific functional words.

The approach for reducing typically derived words to their root kind is known as Stemming. Majority of the words in the English language will be compressed to their base form. e.g. fishing, fished and fisher belong to the root word, fish. Pre-processing is done using Porter-Stemmer algorithm as illustrated in Figure 2.

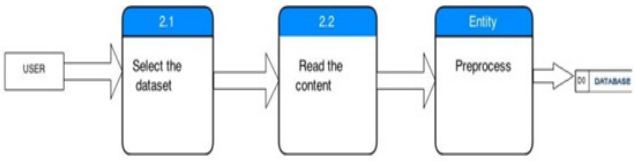

Figure 2. Data flow diagram for preprocess.

5.2 Topic Generation Module

The Figure 3 shows the Flow for Topic Generation. The user will select the preprocessed file from the database. The user will submit the file to tool which will generate the related topics based on the text in the file. This process is done by Latent Dirichelt Algorithm. Initially a distribution is chosen over a combination of topics to arbitrarily choose a topic from the topic distribution. Finally a word is generated from the topic-word distribution. The main output will be the list of topics at hand in the text.

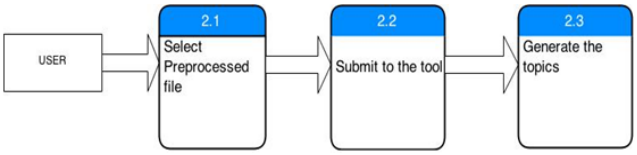

Figure 3. Data flow diagram for topic generation.

5.3 Polarity Classification Module

In Polarity Classification module, the preprocessed file from the database will be selected by the user. The selected file will be submitted to the Maxent tool and the subjectivity of the text in the file will be created as shown in the Figure 4. The text is tagged as noun, verb, adjective, adverb, conjunction.

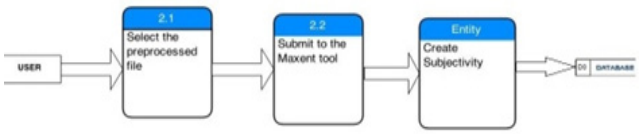

Figure 4. Data flow diagram for polarity classification.

5.4 Sentiment Classification Module

Select the topic which is already been generated and select the subjectivity file from database. With the help of the K-means algorithm the required content will be created which is classified as the result as shown in the Figure 5.

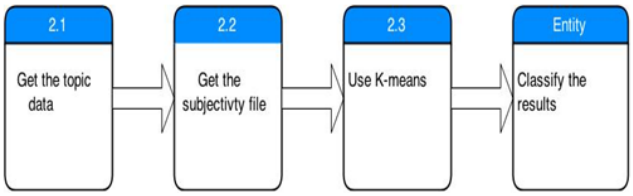

Figure 5. Data flow diagram for sentiment classification.

5.5 Sentiment Analysis Module

In Sentiment Analysis module, select the result of the classification module. Calculate the occurrence of repeated sentiment words based on the subjectivity. Calculate the weightage values for each sentiment words. The higher wieghtage value word will be the final sentiment word as shown in the Figure 6.

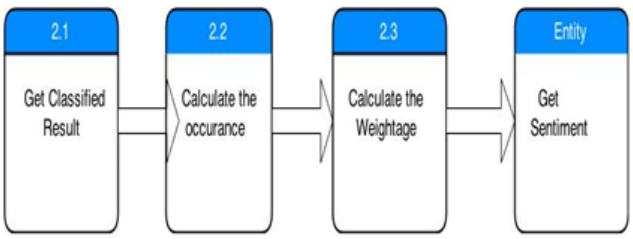

Figure 6. Data flow diagram for sentiment analysis.

5.6 Aspect Ranking Module

In Aspect Ranking get the output of sentimental analysis module , select all the higher values and then rank the higher sentiment word at the top based on the priority as shown in the Figure 7.

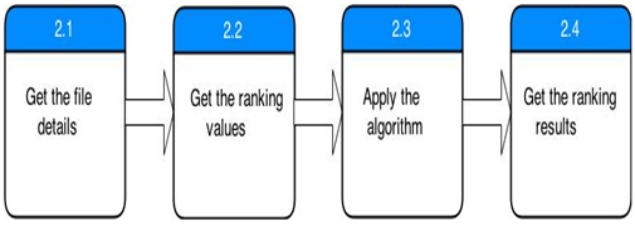

Figure 7. Data flow diagram for aspect ranking.

K-means clustering is the elementary algorithm for clustering. Initially the inputs that are taken up by the algorithm are collection of documents, number of clusters (K) and centroids of each cluster. The space details of documents from the initial centroids are found by using the algorithm to appoint the documents to the nearby clusters. The method is continued until some terminating critera is satisfied. Initial centroid selection is done promptly. Once the clusters are identified, it then procreates the cluster label by searching the documents for terms with higher frequency.

5.6.1 Sentiment Detection

The topic words are processed and the text is tagged for sentiment detection.

5.6.2 Feature Extraction

The detected sentiment is used as input for feature extraction in order to process the sentiment encountered.

5.6.3 Social Emotion Prediction

The feature extraction and sentiment detection are used to predict the social emotions from user reviews.

6. Proposed Method

Customers write opinions about products and about other services. The opinions are collected as documents and are submitted as input to the opinion mining process as shown in Figure 8. A specific domain is selected for data mining processes such as preprocessing, removal of stops words and stemming.

Figure 8. Obtaining the input file.

The pre-processing step is carried out as shown in Figure 9. In case of stop word removal, the words like a, an and are removed. The root word identification is done by implementing Porter stemming algorithm.

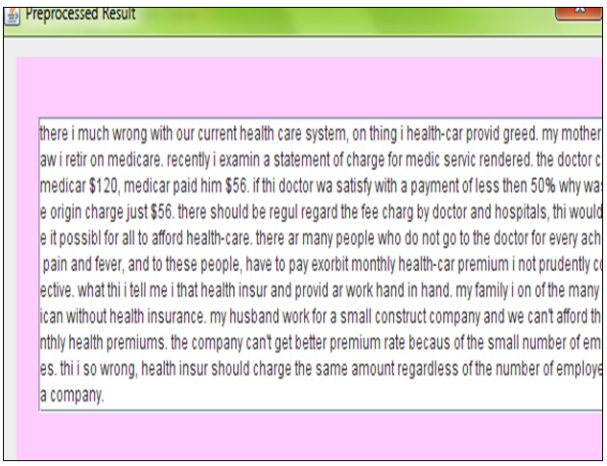

Figure 9. Preprocessing.

The tool will generate the related topics based on the text in the file as shown in Figure 10.

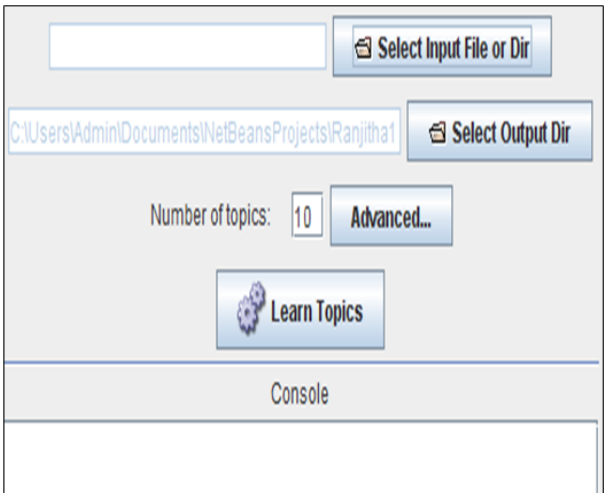

Figure 10. Topic generation.

POS tagging process is performed. The text is tagged as noun, verb and adjective as shown in Figure 11. The tagged text is provided as output.

Figure 11. Polarity classification.

In this process, the string tokenization is performed. The input obtained by this step is the pre-processed text as shown in Figure 12. First a distribution is chosen over a mixture of topics. The topic distribution is used to randomly choose a topic. Eventually generation of a word occurs.

The sentiment detection process takes place from the available text. Using the K-means clustering clusters are generated. The Figure 13 illustrates the identified emotion that is present.

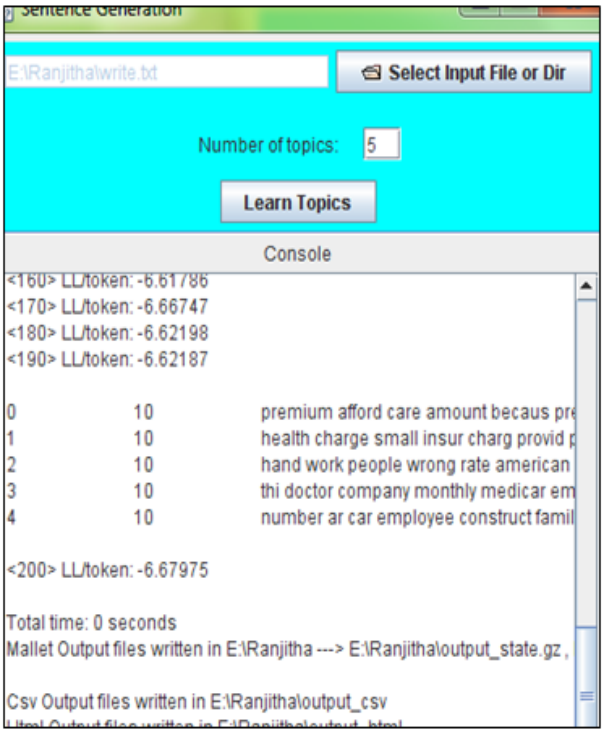

Figure 12. String tokenization.

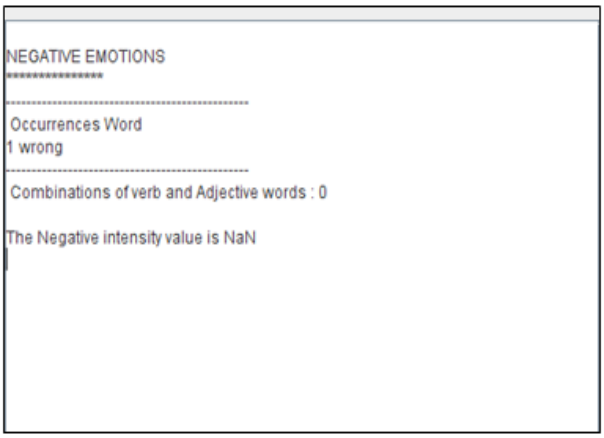

Figure 13. Screen shot for sentiment classification.

As shown in Figure 14 the output of sentimental analysis module is obtained and all the higher values are selected to rank the higher sentiment word.

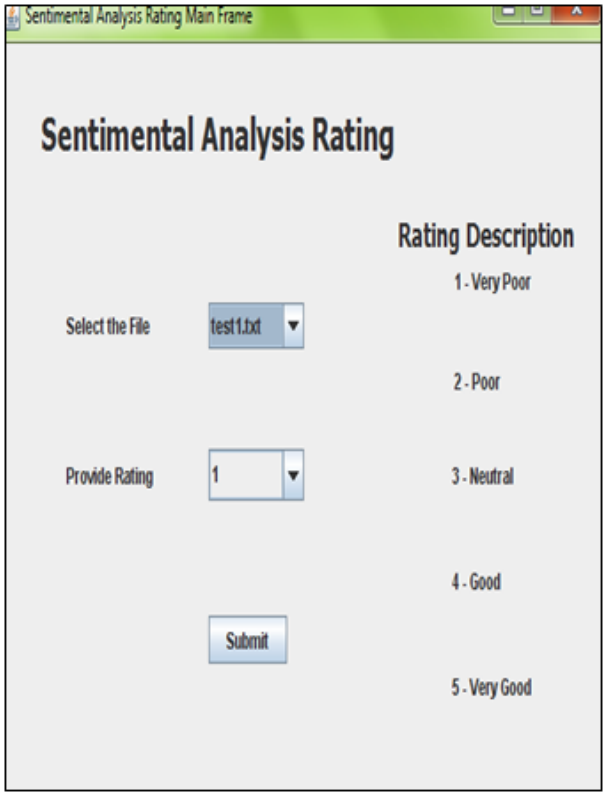

Figure 14. Screen shot for aspect ranking.

7. Conclusion

This machine learning process is able to make a careful determination of patients opinion about the various administration aspects of a hospital based on the prediction accuracy that have been achieved. Various machine learning predictions are correlated with results of many other conventional surveys.

8. Future Enhancement

For future work, it is interesting to generate more efficient algorithms based on topic models in several other opinion mining systems and for large-scale data sets.

9. References

1. Kim SM, Hovy E. Determining the sentiment of opinions. Proceedings of the Conference on Computational Linguistic; 2004. p. 1–8.
2. Cai R, Zhang C, Wang CCC, Zhang L, Ma WY. Music sense: Contextual music recommendation using emotional allocation. Proceedings of 15th International Conference on Multimedia; 2007. p. 553–6.
3. Alm CO, Roth D, Sproat R. Emotions from text: Machine learning for text-based emotion prediction. Proceedings of Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP '05); 2005. p. 579–86.
4. Verma A, Gahier AK. Topic modeling of E-News in Punjabi. Indian Journal of Science and Technology. 2015 Oct; 8(27):1–10.
5. Strapparava C, Mihalcea R. Learning to identify emotions in text. Proceedings of 23rd Annual ACM Symposium on Applied Computing,(SAC'08); 2008. p. 1556–60.
6. Blei DM, Ng A, Jordan MI. Latent dirichelet allocation. The Journal of Machine Learning and Research. 2003; 3:993–1022.