Grenze International Journal of Computer Theory and Engineering , Feb 2017

Sexual Category Feature Extraction of Sentiments about Ayurvedic Health Treatment

Deepa Mary Mathews[1] and Sajimon Abraham[2]
[1]School of Computer Science, M.G University, Kottayam
Email: deepaprasad114@gmail.com
[2]School of Management and Business Studies, M.G University, Kottayam,Kerala,India
Email: sajimabraham@rediffmail.com

***Abstract*—Ayurveda, a form of Traditional Indian Medicine (TIM), literally translates from Sanskrit to "knowledge of life" or more specifically "systematic knowledge of the lifespan". In fact India is the motherland of Ayurveda, and because of its incredible benefits, it gains world wide acceptance and fame. Ayurvedic treatment induces deep recreation and peace of mind. Many ayurvedic hospitals offer these conventional ayurvedic treatments and drugs for any type of illness. Today's patients have start in on to report their health care experience on the Internet in blogs, social networks, wikis, and on health care rating websites. With the explosive growth of social media, patients are increasingly using these media for their decision making regarding the healthcare. Most of us often seek out the opinions of others, when we need to make a decision. New analytical method, such as sentiment analysis, may allow us to recognize and use this information more efficiently to perk up the quality of health care. Various demographics for example gender or age can demonstrate considerable variation in their language used, particularly in informal contexts.**

***Index Terms*— Opinion Mining, Sentiment Analysis, Ayurveda, Classification.**

I. INTRODUCTION

Ayurveda, the conventional medical system of ancient India, nurtures an individual's physical strength and brainpower along with preserving a great equilibrium with the environment. Ayurveda is fully recognized as a medical science by the World Health Organization (WHO) and has amassed an enormous wealth of empirical healing knowledge. In India and in some neighboring countries, Ayurvedic medicine is officially and legally recognized as on par with conventional medicine and they are practicing Ayurveda [4]. Approximately 75% - 80% of the people in Nepal, Sri Lanka, China, European and Western countries use some form of ayurvedic products and the governments have established various medical regulations and universities to offer Ayurveda practice to common people too, so as to enable this medicine system a ceaseless practice [3]. The Indian ministry of Ayurveda, Yoga and Naturopathy, Unani, Siddha and Homoeopathy (Ayush) is now all set to sign a pact with the World Health Organization (WHO) to spread the significance of Ayurveda and will work with the global organization to celebrate a World Ayurveda Day on 28th October. The importance of Ayurveda in modern South Asian health care setups is reflected by the following figures: In India alone above 400,000 registered Ayurvedic physicians practice Ayurveda [1] and there are more than 250 universities and colleges where Ayurvedic medicine is systematically taught as a 4–6-year university degree program [8]. In addition to its key role in Asian health care systems, it is playing a growing role in Complementary and Alternative Medicine (CAM), especially in integrative

Grenze ID: 01.GIJCTE.3.1.1

settings in Europe and North America. For instance, in Germany, Austria, and Switzerland Ayurveda is one of the fastest growing CAM methods [4]. The World Ayurveda Foundation (WAF) is associate degree initiative by Vijnana Bharati geared toward world propagation of Ayurveda.

Kerala Ayurvedic Treatment induces deep relaxation and peace of mind. Many ayurvedic hospitals especially located in Kerala offer these conventional ayurvedic treatments and medicines for any type of illness. The examination of the illness is done through the incorporation of physical and mental characteristics. There are eight ways to identify all types of illnesses, namely Aakruti (appearance), Druk (vision), Jihva (tongue), Mala (stool), Mootra (urine), Nadi (pulse), Shabda (speech) and Sparsha (touch) and five senses (hear, sight, smell, taste and touch) [3].

Typical measures of patient experience include surveys, and more recently, structured patient reported outcome measures. Such approaches ask specific and limited questions, are conducted occasionally, and are often expensive to administer. Today's patients have begun to report their health care experience on the Internet in blogs, social networks, wikis, and on health care rating websites [5]. Immense amount of unstructured, free-text information about quality of many ayurveda hospitals and siddha practioners are available on the Internet in blogs, social networks, and on physician rating websites, but they are not captured in a systematic way and so difficult to analyze the patient's experience. Most of us often seek out the opinions of others, when we need to make a decision. This is true not only for individuals, organizations and also for the doctors in medical field. With the explosive growth of social media in the past 10 years, patients are increasingly using these media for their decision making regarding the healthcare [2]. To analyze these big data, we have to apply various text mining tools to the unstructured data to make it as a meaningful data. New analytical techniques, such as sentiment analysis, may allow us to understand and use this information more effectively to improve the quality of health care. Survey data in the United States suggest that 85% of adults use the Internet, 25% have read someone else's experience about health on a website or blog, and 11% have consulted online reviews of hospitals or other medical facilities [5].

II. Sentiment Analysis

The immense amounts of information available in the form of reviews, blog posts, social media comments and tweets (called as User Generated Content (UGC)) on the web have to be mined which leads to the need of opinion mining and sentiment analysis. Even though we are drowning in data, we are in lack of meaningful data. These unstructured data on the web have to be analyzed to find out a meaningful data. Sentiment Analysis (SA)) is the computational study of people's opinions, appraisals, attitudes and emotions toward entities, individuals, issues, events, topics and their attributes. The whole process of identifying and extracting subjective information from raw data is known as sentiment analysis [7]. It extracts opinions, emotions and sentiments from the data. This process can be used to determine whether a piece of writing in the blog about the treatment is positive, negative or neutral. It derives the opinion or attitude of a patient. Are they happy/ satisfied or not? Opinions of users not only help individuals in taking informed decisions but also help organizations in identifying patient attitudes, opinions about treatments, services etc. Our intention was to test whether we could automatically predict patients' views on a number of topics from their free-text responses.

Machine learning approach relies on the famous Machine Learning algorithms to solve the Sentiment Analysis as a regular text classification problem that makes use of syntactic and/or linguistic features[7]. Machine Learning tools can be used to analyze various sentiments which may be mined to know whether the patient suggest that particular treatment for similar type of diseases, whether the services from the hospital is good or bad, about the cleanliness, hospitality etc. This information helps others to know about the treatment before going for it.

The two components of machine learning approach: (1) preprocessing which cleans the data, and (2) classification, in which an algorithm decides which category each comment falls into [5]. The various processing steps required for getting the useful data from the user data or reviews are shown in figure 1.

Steps in Sentiment Analysis

1. Load the dataset.
2. Preprocessing and Term Document Matrix creation.
3. Classify and create positive and negative word clouds by applying the sentiment algorithm
4. Analyze the result

First collect the available review comments and ratings about the ayurvedic treatments given by various ayurvedic hospitals from various websites and blogs. It is extremely difficult to analyze these big data. It is very time consuming and requires lots of manual effort to read all these reviews and analyze and generate a positive and

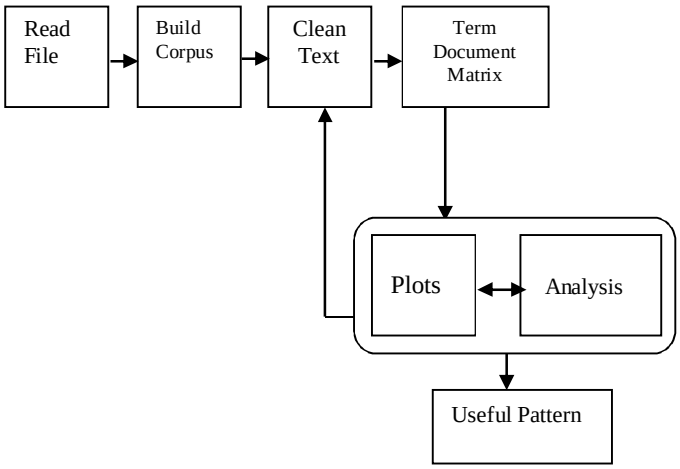

Figure 1: Steps in Sentiment Analysis

negative word cloud. The text input which contains the user reviews can be broken into a set of documents called as corpus. The corpus then should undergo preprocessing to clean the data.

A. Preprocessing

Preprocessing is the process of introducing a new document to the information retrieval system in which each opinion of the patient corresponds to a set of index terms for the efficient storage and retrieval of the data. Preprocessing is from which data from patient comments are split into manageable units to build a representation of the data. This process removes the data which is not required for analysis. Ie cleaning includes striping whitespaces, removing punctuation marks, stemming, removing stop words, removing website links etc.

B. Classification

In data analysis, algorithms have been developed that can be used to analyze data, with the goal to extract useful information. Some widely used classification algorithms are Naive Bayes and Support Vector Machines (SVM). These algorithms can be used to assign a sentiment (positive or negative) to the patients' tweet or opinion in the blog. The number of tweets that have to be processed is too high and complex for normal data adapters, therefore a complex event processing engine is required for processing the data.

Also it is noted that the review comments or the sentiments of the user may vary depends upon the various demographic features like gender, age group, education, marital status etc. Different demographics can demonstrate substantial variation in their language used, particularly in informal contexts. For example, usually the female patients' gives more importance to the cleanliness of the surroundings compared to the male patients, the educated users give more significance to the experience and qualification of the doctors in the hospital etc. Gender differences in subjective language can effectively be used to improve sentiment analysis. Some words are more or less likely to be positive or negative in context depending on the gender of the author. So we can build a classifier to distinguish the gender of a user.

The various steps in building the classifier are:

- Import the Dataset and corresponding class labels.
- Define the feature extractor function
- Building training set and a test set from the Dataset.
- Build the classifier
- Test the accuracy

We can use Naïve Bayes Theorem for Classification.

Naïve Bayes Classifier

It's a probabilistic and supervised classifier given by Thomas Bayes. It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other

feature. According to this theorem, if there are two events say , e1 and e2 then the conditional probability of occurrence of event e1 when e2 has already occurred is given by the following mathematical formula:

$$P(e_1 \mid e_2) = \frac{P(e_2 \mid e_1)P(e_1)}{e_2}$$

This algorithm is implemented to calculate the probability of a data to be positive or negative [6]. So, conditional probability of a sentiment is given as:

$$\text{P(Sentiment} \mid \text{Sentence)} = \frac{\text{P(Sentiment)P(Sentence} \mid \text{Sentiment)}}{\text{P(Sentence)}}$$

And conditional probability of a word is given as:

$$\text{P(Word}\mid\ \text{Sentiment)} = \frac{\text{(No: of words occurrence in class+1)}}{\text{No: of words belonging to a class + Total no: of word}}$$

We can use Naïve Bayes Theorem for gender Classification. According to the classification theorem, mostly names ending in a, e and i are likely to be female, while names ending in k, o, r, s and t are likely to be male. [11]

III. EVALUATION

The algorithm was trained using the comments and ratings about hospitals from various top rating websites like trip advisor dataset. At first, we will extract the last letter of the name of user who is posting the review or opinion in the websites or blogs. The following feature extractor function written in python builds a dictionary named sexualcategory_feature which contains the last letter of a given name:

```
>>> def sexualcategory _features(word):
...     return {'last_letter': word[-1]}

>>> sexualcategory _features ('Haleena')
{'last_letter': 'a'}
```

Use the feature extractor to process the data and divide the resulting list of feature sets into a training set and test set. The last 500 rows from the total dataset in the featuresets is taken as the testing set and the remaining can be considered as training set. We can print the length of training set and the testing set using the python built in function len().

```
train_set,test_set=featuresets[500:],featuresets[:500]
print len(train_set) ,len(test_set)
```

We can construct our classifier based on the training set which uses the Naïve Bayes classification algorithm, then test out the classifier with few samples outside of training set

Length of the Dataset considered - 7944
Length of training set: 6444
Length of test set: 1500

Test Data
Name - Mathews Gender is Male
Name - Prasad Gender is Male
Name - Aleena Gender is Female
Name – Kumar Gender is Male
Name – Surabji Gender is Female
Enter a name to predict the gender:

(Type q to exit) Elizabeth
Female

Accuracy of the Classifier
(which considers only last letter): 0.761333333333

Finally, we can examine the classifier to determine which features it found most effective for distinguishing the names' genders:

```
>>>classifier.show_most_informative_features(5)
Most Informative Features
last_letter = 'a'        female : male   =     33.2 : 1.0
last_letter = 'k'        male : female =     32.6 : 1.0
last_letter = 'p'        male : female =     19.7 : 1.0
last_letter = 'v'        male : female =     18.6 : 1.0
last_letter = 'f'        male : female =     17.3 : 1.0
```

We can reconstruct the classifier by extracting last two names of the user and again while testing the accuracy it is shown that that the accuracy level is increased by 0.1%. Likewise we can consider various attributes which increases the accuracy level of the classifier.

IV. Conclusions

The work can be extended to consider more attributes or domain for classification. The feature selection approach like information gain, gini index etc can be used to improve the computation time and also demonstrates the words with highest predictive accuracy. A number of different technical approaches can be taken to classification in machine learning, to see which gave the quickest and most accurate results.

References

[1] Association of Ayurvedic Physicians of India (AAPI), (2013), http://aapiindia.org/.
[2] Ananda Shankar, Dr.K.R.Ananda Kumar,(2015), "Top K-Opinion Decisions Retrieval In Healthcare System", Computer Science & Information Technology (CS & IT), ISSN : 2231 – 5403.
[3] Gunaseelan, Dr. V.Ramesh, (2016), "A Study on Application of Data Mining in Ayurinformatics", International Journal of Computer Applications (0975 – 8887), Volume 137 – No.4.
[4] Kessler, M. Wischnewsky, A. Michalsen, C. Eisenmann and J. Melzer, (2013), "Ayurveda: Between Religion, Spirituality, and Medicine", Evidence-Based Complementary and Alternative Medicine, Volume 2013 , Article ID 952432.
[5] Felix Greaves, Daniel Ramirez, Christopher Millett, Ara Darzi, Liam Donaldson, (2013), "Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online", Journal of Medical Internet Research
[6] Pravesh Kumar Singh, Mohd Shahid Husain, (2014), Methodological Study Of Opinion Mining And Sentiment Analysis Techniques, International Journal on Soft Computing (IJSC) Vol. 5, No. 1
[7] Shailesh Kumar Yadav,(2015), "Sentiment Analysis and Classification : A Survey", International Journal of Advance Research in Computer Science and Management Studies, Volume 3.
[8] WHO (2010), Benchmarks for Training in Traditional/Complementary and Alternative Medicine: Bench-Marks for Training in Ayurveda, World Health Organization, Geneva, Switzerland.
[9] www.nltk.org/book/ch06.html