ASIAN JOURNAL OF PHARMACEUTICAL AND CLINICAL RESEARCH

INNOVARE ACADEMIC SCIENCES
Knowledge to Innovation

Vol 10, Issue 3, 2017

Online - 2455-3891
Print - 0974-2441

Review Article

PREDICTING RATINGS FOR USER REVIEWS AND OPINION MINING ANALYZE FOR PHYSICIANS AND HOSPITALS

HEMA SAGAR VAKATI*, JEBAKUMAR R

Department of Computer Science and Engineering, SRM University Kattankulathur, Tamil Nadu, India. Email: v.hemasagar@yahoo.co.in

Received: 09 November 2016, Revised and Accepted: 02 December 2016

ABSTRACT

Health care is taking its turn in the internet now and online health information consumption is also booming. Users have started generating healthcare reports like online doctor reviews open to all. Hence, online health forums are increasingly popular these days since people can gather their required data by just sitting at home and select the best doctor by considering the reviews available online. The patients also browse on their concerned diseases and use the open forum for discussion on the topics. On an average, these online health-care providers are mainly focusing on reviews about the physicians. The feedback provided by patients is considered and we also analyze the sentiments of the patient to estimate the value of the reviews. The rating for the doctors is divided into various categories such as Staff, Knowledge, and Helpfulness. We propose support vector machine and apriori for the classification of data and use sentiment based rating prediction to analyze doctor's reviews and opinion mining patterns for online patterns. By providing physician ratings in website, it offers the patients to know about the physician and consider the critique and information to make their decision.

Keywords: Support vector machine, Apriori, Sentiment classification, Opinion mining.

 DOI: http://dx.doi.org/10.22159/ajpcr.2017.v10i3.16094

INTRODUCTION

Online health-related information is the new era in information technology. It has become convenient for the patients to seek information about their disease conditions and to know about their physician and hospitals. They are also explored to information which cannot be easily attained from other sources. On the research in this topic, it is obvious that most of the queries that we deal are to do with hospitals and physicians and especially physician information which includes their performance and ratings are considered. Websites - such as vitals.com, HealthGrades.com, and rate MDs.com - are some of the sites that allow patients to give open feedback on their individual MDs based on various aspects such as staff quality, personal demeanor, and knowledgeability. These websites are open to all, and patients can write critique and also obtain their required information. This information indeed serves good by allowing others to take the right insights from these feedback information [14]. In this paperwork, we propose a method to understand the inputs from the user and take a comprehension from character level information such as words, expressions, example, sentences, and passages to make it relevant to the application.

The above strategies can perform extremely well only when a severe pre-learning process happens, like pre-characterize lexicon of intrigued words, and the parser needs to be trained for several exceptional cases. Cases with word morphological changes and questionable lumping are some of the exceptional cases to be trained and handled [3]. When it comes to online health-care provider reviews, physician reviews are handed over to the medical consultant through the patients, and we review the patient's sentiments to correctly validate the doctor's ratings in various categories such as, knowledge, efficiency, and helpful. In this proposed paper, we have several techniques and approaches to have an opinion oriented information seeking system. Our research is completely based on the new challenges and approaches to fit the requirements. The main area of concern is the sentiment aware application, which is quite different from the traditional fact-based analysis. We consider that the material of text to be considered and other privacy issues, manipulation processes and the economic impact of the opinion oriented information access services.

Background

Srivastava *et al*. [12]. It is common in internet to see searches related to food and nutrition. To get the relevant and reliable information from the search engine is the biggest challenge due to the vast amount of information present in the internet. Especially in the field of healthcare and nutrition it is utmost important to deliver the most reliable data to the user. Hence, the challenge taken up in our research work is to deliver an efficient outcome for user queries. As we all know that it is highly difficult to search for the right content over the vast lying data in the internet. The system is designed in such a way to satisfy the user with the best search results for food and diet recommendations and also the quality of the outcome is taken care. The output has to be of utmost benefit to the user. There are various factors on which the recommendation pattern depends on [16]. This paper proposes a semantic framework prototype to serve the health-care recommendation system. This method helps in extracting the output in favor of client necessity and imperatives.

Paul *et al*. [3]. The system will be able to analyze the user's satisfaction or dissatisfaction depending on the reviews of doctors and prediction of their ratings. We use the Convolutional Neural Network to optimize various aspects and loss functions. We used the 35000 user reviews available in www.ratemds.com for around 10000 doctors. The novel method was able to achieve 93% accuracy rate for positive/negative binary classification of patient reviews. We achieved a mean absolute error of 0.525 for prediction rating on a 5-point scale and with error rate of 0.71.

METHODS

Opinion lexicon expansion

Opinion words are implemented to classify the sentiments. In this topic, we are going to discuss in detail about opinion words and the methods implemented to use them classify text documents. Opinion words are known as sentiment words or opinion bearing words. To classify

sentiments we will deal with categories namely positive and negative. Positive opinion words are to express happiness or satisfaction, while negative opinion words are to express dissatisfaction. Examples are beautiful, good, excellent, amazing for positive opinion, and bad, terrible, sad, poor are for negative opinion words. We will also be dealing with phrases and idioms to represent sentiments like cost someone an arm and a leg. All these together form the opinion lexicon. To collect the entire list forming the opinion word list is done through three main approaches: (1) manual approach, (2) corpus-based approach, and 3) dictionary-based approach. As we know that the manual approach is tedious and time consuming, we combine manual approach with automated approaches as well. Since automated approaches make mistakes at times, we merge the manual and automated approach to deliver error free environment.

Sentiment classification

Sentiment classification is the process by which the opinions in a document are classified based on the overall sentiment orientation. In this approach, we deal with documents which are subject oriented such as product reviews and feedback forms. Opinion orientation can be classified as favorable or unfavorable opinion, positive or negative feedback based on a subject. It is also used to rank an object based on the opinions received. Example of such opinion ranking is the film reviews. A film is ranked from one to five star based on the opinions. We have used supervised learning methods using different features of text as a major source. We have proposed support vector machine (SVM) along with opinion mining classification to obtain best results.

We have also explored classifying a document based on grammar, parts of speech, and many more. In part of speech, the features are extracted to classify the sentiment. When detecting and giving a score pattern for part of speech is by deriving the features for sentiment classification this method is used for extracting product features as well. By portioning the subject and object of a sentence improves the sentiment classification better than other proposed methods. This is done using baseline word vector classifier. Other methods use the correlation of writing style and use colloquialisms and punctuation, which is used to convey sentiment. In the lexicon of colloquial expressions, we have also approached with unusual spellings like great, word combining like super good is also to be used for sentiment classification. We have merged statistics and feature measuring aspects and used along with word vectors to show improvement in the baseline classifier.

Modified SVM

Modified SVM is used for the classification of data. It is considered to deliver better output than the Neural Networks because it always gives accurate and satisfactory results. The process of classification involves testing and training data which in turn consists of data instances. The instances in the training set consist of the target value and the attributes attached to it. The purpose of having SVM in this process is because it evaluates the target value of data instances in the testing data where only the attributes of the same are given.

Classification of data in the SVM model is very similar to the Supervised Learning. It helps the system perform better by systematically arranging the data. It further helps the system to attain accuracy and perform correctly. Identification of relevant data is the important step in SVM, where the data are classified and connected to the right class. This process is called as the feature selection or feature extraction. This process of feature selection along with the modified SVM classification helps in during classification of unknown samples.

Applications of modified SVM

As we know the ultimate use of SVM, there are still some dark areas of applying this method in practical. To rectify this modified SVM was proposed which works well in case of pattern classification area. Modified support vector not only serves the purpose of data classification but also helps in solving other problems related to the objective of the project. One such major problem is with choosing the right kernel for the given application. Some of the commonly used kernels are the polynomial kernel and the Gaussian kernel, but they do not work well in case of discrete structures. In such cases, we need an elaborate kernel in use. As in the above case, by defining the exact feature space, it is possible for the kernel to provide the description language which is used by the machine to view the data. Once the kernel is finalized and the optimization criteria are set the system is very much in place. Hence, let's have a look at an example.

Classification of text documents is categorizing them into the predefined categories based on the content in the document. As we know that a document can fit into more than one category, it is definitely a problem when it comes to multi-class classification. A traditional way of representing text documents is through ideal feature mapping with Mercer kernel. The kernels find a way to have the similarity measure between instances and the researches who work in this application would have already established the similarity measures.

The traditional approaches fail in such cases because we work with high dimensional data and to face the present challenges we use the modified SVM. The successful approach used in text classification can be well used in image classification as well through linear hard margin machines. The first ever successful experiment on modified SVM was on handwritten character recognition. Multi-class modified SVMs were also tested on these cases. The various modified SVMs are also compared to study their unique features. The experimental results show that the SVMs perform very similar with few minor variations in the end result.

APRIORI ALGORITHM

General process

The two main processes involved in Apriori algorithm are as follows:

1. Minimum support is used to find all the frequent itemsets in the database.
2. The retrieved frequent itemsets and the minimum confidence constraint are applied to form the rules.

Fetching the frequent itemsets is a tedious task as all the itemsets are to be searched. The set with all possible itemsets is the power set over I and with size 2n−1 (which does not contain the empty itemset). Although the power set can vary through its size with the increase in the number of items in I, it is possible to have an efficient search using the downward closure property of support (called as the anti-monotonicity). This rule states that for a frequent itemset, all its subsets are also frequent and for infrequent itemset its supersets are also infrequent. Using the above property, the Apriori algorithm can find all the frequent itemsets.

Apriori algorithm pseudocode

```
Procedure Apriori (T, minsupport) {//T is the database and minsupport
is the minimum support
L1= {frequent items};
For (k = 2; Lk-1 ! =∅; k++) {
Ck= candidates generated from Lk-1.
//that iscartesian product Lk-1 x Lk-1 and eliminating any k-1 size
itemset that is not//frequent for each transaction t in database do{
#increment the count of all candidates in Ck that are contained in t
Lk = candidates in Ck with min support
}//end for each
}//end for
return ∪ k, Lk-;}
```

As a common concept in association rule mining, with a given set of itemsets (it could be a list of items an individual purchased or retail transactions), the proposed algorithm tries to find subsets that are common to find a minimum number C. In Apriori, the "bottom-up" concept is used where the frequent subsets are expanded with one item at a time, and post which the groups of candidates are to perform tests against the data. The algorithm will end once there is no more extension in the group.

Breadth-first search and tree structure concept is used to count the candidate itemset in Apriori. It will form candidate itemset of length k from k-1 itemsets. The candidates with infrequent pattern are removed and a set of candidates with frequent k-length itemsets are formed using downward closure lemma. Now the comparison is made with the transaction database to find the frequent itemsets among the candidates. Although Apriori is a successful algorithm, it still has its own inefficiencies. In the process of candidate generation, a large number of subsets are generated. The bottom-up will find only the maximal subset S only after all 2|S|−1 of its proper subsets.

DATA PREPROCESSING FOR USER REVIEWS

This method is mainly proposed for large scale speech recognition dataset and MNIST. The proposed method was able to deliver promising results. It has the capability to perform trivial computational tasks than the SGD and providing pre-dimension learning rate [11]. When it comes to speech recognition, there are a lot of things to consider such as chunking, parts-of-speech tagging, semantic role labeling, and named entity recognition. The system adapts these features by forming a strategic distance from errand particular building and ignoring the early learning. Our proposed framework is trained to learn interior representations on the premise [5]. Although there are various forms of input data types, the system need not be predefined to process these data, hidden units, distributed replicas, hyperparameters, and nonlinearities. This shows that ADADELTA is a robust system and be used in various scenarios [2]. Deep learning to understand the texts and all the way to abstract text concepts with the use of temporal convolution networks also called as ConvNets. ConvNets are applied in various platforms such as sentiment analysis, large-scale datasets, ontology classification, and text categorization. ConvNets can deliver amazing results without the knowledge on sentences, phrases, words or any other syntactic or semantic structures. Simulation results show that the proposed model can work efficiently in English and Chinese.

Rama [7]. proposed the phoneme level Siamese for pairwise cognate identification. This represents a word in two-dimensional matrix and applies Siamese convolutional network for understanding learning deep representations. The Siamese method helps in understanding phoneme level feature representations and language relatedness from raw words for cognate identification. To show that the Siamese architecture performs better in a large and realistic dataset, we will be performing computation using continuous vector representations of words taken from a large dataset.

The performance and quality of this method is compared with other traditional methods which work on different types of neural networks. Accuracy is achieved using low computational costs (it's less than a day taken to learn the high-quality word vectors from 1.6 billion words data set). These vectors perform state-of-the-art and deliver syntactic and semantic word similarities [4]. The various methods computational system is evaluated to analyze the performance of our proposed method. The model is constructed to perform on the star ratings taken from the Google Play Store.

OPINION MINING FOR SENTIMENTAL DATA

There are methods already in place to analyze the sentiment orientation called the fine-granular opinions. These models are building to predict the sentiment based on polarity. While performing the ratings of products, it is very important to consider the sentiment of the customer. Rating the products by considering the written reviews has become a tedious task for automated predictors. Hence to solve this complicated task, we have proposed a series of experiments using convolutional neural networks using word to vector. With a simple change in the process using a simple conventional neural network with an addition of convolution performs extremely well. The results show that the method can work well even during unsupervised pre-training of word vectors [12]. The most suitable way to approach these altered methods is to normal the expectations of the parameters.

Turian *et al*. [8]. First, we need to create a systematic way to compare different words in a controlled way. Word embedding and Brown clusters can improve the accuracy of near state-of-the-art with NLP system. It is also noticed that by combining different word representations can also improve the performance of the system. It is observed that using the Brownian clustering the rare words is better represented. Another contribution to this paper is the setting of default method for the scaling parameter for word embedding's [9]. This model is efficient that it trains only the nonzero elements present in the word-word co-occurrence matrix and not the entire matrix.

Zeiler [6] Another approach to earn better performance rate is by rectifying neurons and creating sparse representations with true zeros. Deep rectifiers networks perform extremely well even with unsupervised pre-training with large labeled datasets. This is indeed a new approach to the attempt of understanding the difficulty in training deep [13]. This proposed paper summarizes on entailment acknowledgment, talk examination, and summarize location, opinion examination, machine translation, grounded dialect learning, and picture retrieval.

CONCLUSION

In the field of online health-care provider services, physician reviews are the mainly browsed data. The feedback about the physician is taken first-hand qualitative from the patient. Using this feedback data, we consider the sentiments of the patients who provided the reviews of doctor under various categories such as "helpfulness," "staff," and "knowledge." We used the SVM and Apriori algorithm to extract the sentiment based rating and for doctors reviews, we used the opinion mining patterns.

REFERENCES

1. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in Vector Space. arXiv:1301.3781v3 [cs. CL] 7 Sep 2013.
2. Zhang X, Cun YL. Text Understanding from Scratch. arXiv:1502.01710v5[cs. LG] 4 Apr 2016.
3. Paul MJ, Wallace BC, Dredze M. What Affects Patient (Dis) satisfaction? Analyzing Online Doctor Ratings with a Joint Topic-Sentiment Model. Copyright ©2013.
4. Monett D, Stolte H. Predicting star ratings based on annotated reviews of mobile Apps. dx.doi.org/10.15439/2016F141 24 October 2016.
5. Galizzi MM, Miraldo MM, Stavropoulou C, Desai M, Jayatunga W, Joshi M, Parikh S. Adadelta: An adaptive learning rate method. arXiv:1212.5701v1 [cs. LG] 22 Dec 2012.
6. Zeiler MD. Deep sparse rectifier neural networks. Copyright 2011 by the authors.
7. Rama T. Siamese Convolutional Networks for Cognate Identification. 24 October 2016.arXiv:1605.05172v2 [cs.CL].
8. Turian J, Ratinov L, Bengio Y. Word representations: A simple and general method for semi-supervised learning. 11-16 July 2010. ©2010.
9. Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. October 25-29, 2014, Doha, Qatar. © 2014.
10. Kim Y. Convolutional Neural Networks for Sentence Classification. October 25-29, 2014, Doha, Qatar. © 2014.
11. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuglu K, Kuksa P. Natural Language Processing (Almost) from Scratch. J Mach Learn Res 2011;12:2493-537.
12. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from over fitting. J Mach Learn Res 2014;15(1):1929-58.
13. Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modeling sentences. arXiv preprint arXiv:1404.2188, 2014.
14. Zhang X, LeCun Y. Text understanding from scratch. arXiv preprint arXiv: 1502.01710, 2015.
15. Turian J, Ratinov L, Bengio Y. Word representations: A simple and general method for semi-supervised learning. In: Proceeding the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, July 2010. p. 384-94.
16. Ahire SB, Khanuja HK. A personalized framework for healthcare recommendation. Int J Comput Appl 2015;110(1):89-92.