Patient Opinion Mining: Analysis of Patient Drugs Satisfaction using Support Vector Machine and Logistic Regression Algorithm

Dina R. Nahma[1] Ayad R. Abbas[1]
[1]Computer Science Department \ University Of Technology, Baghdad, Iraq
112056@student.uotechnology.edu.iq ayad.r.abbas@uotechnology.edu.iq.

Abstract

Patients opinion mining is reflect patients' observations and perspectives on medical drugs or treatments in drug reviews. The Opinion Model used in this paper focuses on predicting the degree of drug satisfaction among patients. This work aims to apply social-web opinion mining machine learning methods to healthcare. This data set was collected from a web said UCI learning software repository with two sets of data: train and test (75-25%). Where divided the drug ranking into three general classes: positive (7-10), negative (1-4) or neutral (4-7).Where examine reviews of different drugs in this research, which were analyzed in text format and also assessed on a scale from 1 to 10. After that input the dataset to classification by two algorithms used the support vector machine and Logistic Regression algorithm, used this algorithm to improve the results of support vector machine algorithm. The support vector method with logistic regression yields better results in terms of accuracy.

Keywords: Opinions mining; Support vector machine; Regression algorithm; Drugs reviews; Drugs Satisfaction

تعدين رأي المريض: تحليل رضا المريض عن الأدوية باستخدام آلة المتجهات الداعمة وخوارزمية الانحدار اللوجستي

دينا رحيم نعمة اياد روضان عباس
قسم علوم الحاسوب / الجامعة التكنولوجية ، بغداد ، العراق

الخلاصة

يعكس التنقيب عن اراء ملاحظات المرضى ووجهات نظرهم حول الادوية اوالعلاجات الطبية في مراجعات الادوية.يركز نموذج الرأي المستخدم في هذه الورقة على التنبؤ بدرجة الرضا عن الادوية بين المرضى . يهدف هذا العمل الى تطبيق اساليب التعلم الالي لتعدين اراء الويب الاجتماعي على الرعاية الصحية. تم جمع مجموعة البيانات هذه من احد مواقع الويب المذكورة في مستودع برامج التعلم مع مجموعتين من البيانات : التدريب والاختبار(25%-75%) حيث تم تقسيم ترتيب الدواء الى ثلاثة فئات عامة: ايجابي(7-10) ، سلبي (1-4) ، محايد (4-7) ، حيث يتم فحص مراجعات الادوية المختلفة في هذا البحث والتي تم تحليلها بصيغة نصية وتقيمها ايضاً على مقياس (1-10) . بعد ذلك الادخال ، تم تصنيف مجموعة البيانات بواسطة خوارزميتان الة متجه الدعم والانحدار اللوجستي حيث يستخدم الة متجه الدعم بعد ذلك الانحدار اللوجستي لتحسين نتائج الة متجه الدعم . تؤدي طريقة الة متجه الدعم مع الانحدار اللوجستي نتائج افضل من حيث الدقة.

الكلمات المفتاحية: تعدين الاراء ، اله متجه الدعم ، خوارزمية الانحدار، استعراض الادوية ، الرضا عن الأدوية.

Introduction

Opinion mining is a process in which subjective information, feelings and opinions are detected and analyzed in large volumes of texts using computational methods [1]. Opinion mining in different areas has been applied; especially test product and service acceptance as well as general feelings towards people and brands. Most researches focuse on opinion mining in drug reviews that reflect patients' observations and perspectives on medical drugs or treatments. In recent years this area of application has been very active [2].Opinion mining benefit to drug manufacturers in particular in the field of Pharmacovigilance. Since particular adverse reactions to a medicine are more quickly traced from public storing or social network postings. Prior works on opinion mining in drug reviews focused on classifying the drug feelings expressed in user reviews, Positive and negative. [3][4]

There is an implicit trade-off between benefit of drugs and potential for damage with any medication. In order for the health and wellbeing of the patients to make the best decisions, it is essential to inform and understand the potential risks. Drug monitoring and social media can be used as alternative potentials to complement the existing framework of post-market surveillance and improve it for their quantity and expediency. [5]

Social media has spreaded dramatically across the world over the past decade, and vast amounts of information have become open on the Internet as people have become more likely to share their thoughts about goods, services, movies or whatever they want to share online [6]. The amount of subjective knowledge that can be accessed from the Internet has considerably increased. A growing number of websites are now available where users can create and share what they want. In particular, social networks are a highly regarded data exporter that allows users to express their perspectives, thoughts and feelings. [7]

In this classification issue, many natural language processing (NLP) algorithms and high-tech learning machines were used and the features were translated into numerical data via textual data. The areas of health and medical care are untapped. The aim of this paper is to develop an effective way of analyzing the sentiment of social media content in health and medical fields. The sentiment analyze is used on a discussion sites to do drug analyses.

Literature Review

Drug evaluations and medication monitoring literature can generally be divided according to classification problems like the automated identification of alternative dispute resolution (ADRs) or side effects and to evaluate overall or aspect-based feelings. The majority of approaches to ADR or to recognize side effects is lexicon based and rely on the associated terms and expressions from user-data to unique vocabulary from different individual or combined lexicons [8][9] .However, phonetic and typographic misspellings affect lexicon based approaches. Recent work has therefore also focused on machine learning technology in order to overcome such constraints.

Sentiment analysis of the drug assessment makes it possible for the consumer to select the best medications and also for drug producers and physicians to obtain useful overviews of public opinion and input. The followings are studies that used deep learning to analyze the opinion of patients on drugs.

a. Yazdavar et al in (2016) [10] worked Fuzzy based implicit emotional research on objective sentences the technique used fuzzy method and the dataset used from www.askapatient.com , the result of precision was 81percent.

b. Vinodhini et al in (2017) [11] applied Patient opinion mining for drugs satisfaction review using supervised learning, used SVM (Support Vector Machine) to analysis patient opinion Probabilistic neural network (PNN) and Radial basis function neural networks (RBF) the algorithms for training and testing. The dataset used from www.askapatient.com, the result of Accuracy was 88.6 percent for PNN, and 93.8 percent for RBF.

c. Felix Graber et al in (2018) [12] Aspect-based sentiment analysis of product ratings with cross-domain and cross-data learning used SVM (Support Vector Machine) the algorithms for training and testing From the dataset www.Drugs.com, the result of Accuracy was 70.6 percent.

d. Min Zhang et al in (2019) [13] Adverse Drug Incident Detection Using a Slowly Supervised Convolutionary Neural Network and a Recurrent Neural Network System the technique used Weakly Supervised Model And the dataset used from www.askapatient.com , the result of Accuracy was 83.29 percent.

e. Sairamvinay et al in (2020) [14] Study of the opinion in drugs tests using supervised machine learning algorithms the technique used Artificial Neural Networks method and the dataset used from www. Drugs.com, the result of Accuracy was 91.286percent.

f. Brent Biseda et al in (2020) [15] Enhancing Pharmacovigilance with Drug Reviews and Social Media the technique used baseline models. The dataset used from www.Drugs.com, the result of Accuracy was 94.8 percent.

Theoretical Background

Dataset

Chosed the dataset from the UCI machine learning repository. The entire data has been rendered in two files: train and test tsv files. The number of samples train (75%) and test (25%). The data collection includes patient reports and associated conditions of different medications. The reviews are further grouped into reports on the advantages, side effects and general comments in three respects. Furthermore the data were obtained from online pharmaceutical reviews and are ratings for overall content, a 5-step rating for the side effect and a 5-step rating for efficiency. The aim was to examine the feelings of patient in drugs in a number of different areas, i.e., feelings of effectiveness, side effects and transferability of models across areas. Information attribute consist of (Name of Drug, condition, Review of benefits, Side effects review, Comments, Score, side effects,Performance).

Table (1) Data Description

Data Druglib.com	Train	Test	Condition	Drugs	Length	Rating	Label	%
Overall Rating	3107	1036	1808	541	277.75	Rating ≤ 4 4 < rating < 7 Rating ≥ 7	-1 0 1	21 10 69
Benefits (Effectiveness)	3107	1036	1808	541	212.87	Ineffective Marginally/Moderately Effective Considerably/Highly Effective	0 1 2	8 19 73

Classification Algorithms

The area unit used for classification in three classes is positive, negative and neutral. Classification algorithms for sentiment analysis, rely on the supervised action plan, need to be trained, with examples pre-marked. The model used to identify the domain-specific data should be trained. Marking will be completed through the expression of judgment and polarity of sets of jobs.

1. **Support Vector Machine Algorithm (SVM)**

SVM is a strong classifier derived from the theory of statistical learning that has proved effective for different tasks in text classification .SVM model is employed using python. [16]

SVM should be a supervised model of learning. This model is an algorithm for the study of statistics and the recognition of the classification sequence. The SVM formula is entirely based upon a choice plane that determines boundaries. A preference plane crew with unique classification instances as shown in figure(1). [17]

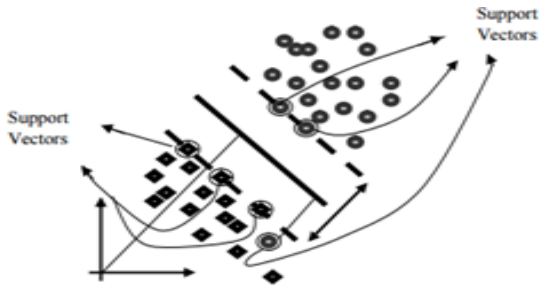

Figure (1) Principle of SVM

SVM is basically valid for two class tasks, apart from multiclass issues, multiclass SVM is available. The case mark area unit provides for at intervals items the area unit has drawn beginning with a restricted location of different pieces. Supporting vector machines algorithmically establish optimal boundaries between data sets by solving a restricted problem in quadratic optimization. Various non-linearity and versatility levels can be used in the model with the use of specific kernel functions. Since advanced statistical ideas can be extracted and boundaries of the generalization error can be calculated, vector supporting machinery in recent years has been receiving a significant interest in science. The medical literature has recorded outputs which are equal to or surpass those of other machine learning algorithms [16] [17]. SVM Advantages It provides excellent efficiency and SVM interpretation drawbacks are problematic. If there are missing values, it should be pre-processed. [16]

2. **Logistic Regression Algorithm**

It is borrowed from the field of statistics like numerous other machine learning techniques and despite its name, it is not an algorithm for regression issues, where a continuous result will be forecast. Rather, logistic regression is classification method. It gives a discreet response from -1 to 1. In simpler words, the result is either one thing or the other. [18]

Logistic regression measures the relation between the dependent variable and one or more separate variables (features) using the underlying logistical functions in order to estimate probabilities [19]. Such probabilities must then be converted into values such that a forecast can actually be made. This is also known as the sigmoid function of the logistic function. The Sigmoid function is an S-shaped curve which allows for any real-value number to be mapped to a value between -1 and 1, never at precisely those limits. These values between -1 and 1 are then transformed by means of a threshold classifier into either -1 or 1. [18]

It's a widely used method, because it's highly efficient, needs no unnecessary computational resources, is highly interpretable, doesn't need to size data, requires no tuning, is easy to regulate, and it delivers well-calibrated predicted

likelihoods [20]. The logistic regression provides an additional benefit that it is extremely easy to implement and effective to train. As logistic regression is relatively simple and quick, it is also a great basis for measuring the performance of other more complex algorithms. The Logistic regression is easier and faster to implement .A drawback of this, since the decision surface is linear, we cannot solve non-linear regression problems**.** [19]

Performance Measures

Used accuracy to measure the effectiveness of the proposed model and compare it with other research results and different methods. Accuracy is defined as number of correctly classified reviews on the total number of reviews. [11]

$$Accuracy = \frac{No.of\ correctly\ classified\ reviews}{Total\ no.of\ reviews} \quad \dots (1)$$

The Proposed Methodology

The methodologies used in this work are listed in detail in this section. The following is the summary of our methodology for improving classification models as shown figure (2).

I. Perform data preprocessing.

Ii. Sentiment analysis by Convert data to numbers for the drug reviews (rating, effectiveness and side Effects).

Iii. Apply the following classification methods using the train data set

A. SVM.

B. logistic regression.

Iv. Input the test data set after preprocessing.

V. Compute accuracy of this work (train, test).

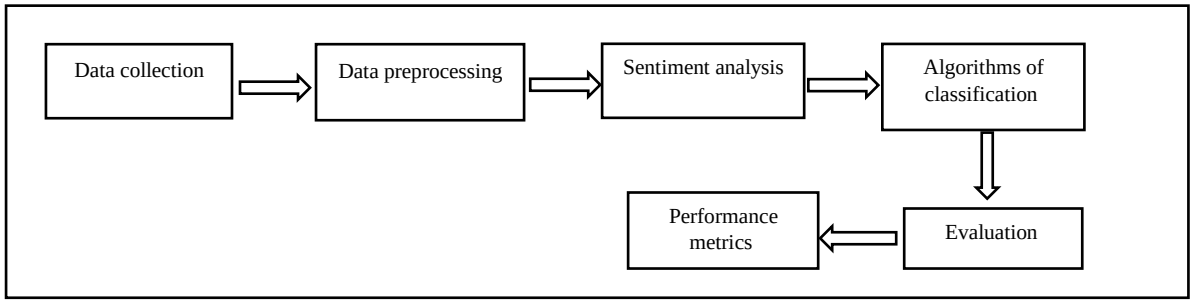

Figure (2) Proposed Solution

Feature Selection For Proposed System

Sentimental research concerning overall satisfaction of patients was translated into classification problems with aspect-based analyzes of feelings of patients about side effects and drug effectiveness. In case of overall patient satisfaction, user ratings were converted to 3 classes which represented the polarity of a patient's feeling with regard to the drug used (negative, neutral, positive). In addition, as defined by Table 1, the rating, effectiveness and side effects were translated into three disjoint groups. Logistic regression for building sentiment models for the various prediction tasks was used with extracted functional representations.

Sentiment Analysis

Druglib.com gathered anonymous user reviews of medications, their side-effects and a quantitative user review score and is available in the UCI learning machine repository. Where first job is to identify these user reviews according to self-reported drug ratings. The quantitative result allows sentiment analysis to be used for a positive evaluation by aggregating highly positive (8 or higher), whereas highly negative (3 or lower) and neutral (4-7) values are also aggregated, where thus qualified this classification of feelings to distinguish these three classes.

Data Preprocessing

At first, all reviews had been preprocessed by a standard scheme: alphabetic characters had been transferred to lowercase and particular characters had been deleted, punctuations and numbers removed. The preliminary documents were subsequently tokenized to obtain a total vocabulary and feature space of each review However, terms with a relative frequency of document higher than a certain limit are discarded when constructing the wording, in order to reduce the feature space.

Classification Methods

In this work used two to obtain acceptable results in this field. The two algorithms were used one by one with the Python programming language.

Used support vector machine to classification and used Logistic Regression algorithm after that, used this algorithm to improve the results of support vector machine algorithm.

Support Vector Machine & Logistic Regression Algorithm

Can construct SVM classifiers using the whole positive examples against negatives and neutral. Then, can integrate these SVM classifiers with logistic regression models to improving the result below is a procedure.

Procedure II

Step 1. Use SVM for train data to produce classifiers (1, ⋯, 10) based "rating class". for each j = 7, ⋯ , 10, to train Classifier using j as the positive set and i=1,⋯ ,4 ; i as the negative set and k=4,⋯ ,7 ; k as the natural set .

Step 2. Determined (positive, negative, natural) from rating class to the separating hyper plane of each classifiers; i.e. for each subject there is a vector d = (1, 2, ⋯, k), where d denotes the distance of the subject to the separating hyper plane of the i-th classifiers.

Step 3. Perform logistic regressions using d as vectors of covariates/regressors.

Step 4. Calculating the accuracy, for all training examples using the final model obtained in Step3.

Step 5. Input a testing data set to model.

Step 6. Repeat Step 1 to step 3 for test data set, and calculating the accuracy, for all testing data using the final model obtained in Step 6.

Results

This section describes experiments to assess the approach proposed. Firstly, prepared a dataset for the development of the algorithm. The dataset has been modified and divided into categories positive, negative, or neural**.** In order to provide comparison with this approach, machine learning approach assessed. In this work a widely used machine learning algorithm for the machine training approach, SVM (Support Vector Machine) and after that used Logistic Regression algorithm to improve the results of SVM.

The support vector machine algorithm when used gave accuracy results for both train data (98.7%) and test data (68.2%) and when used Logistic Regression algorithm gave accuracy results for both train data (99.3%) and test data (72%).

Two algorithms are decided to be used together to improve the results. Initially be using the support vector machine algorithm and then Logistic Regression algorithm to improve accuracy and got higher accuracy for both the train and test data where the accuracy of the train data (99.8%) and the accuracy of the test data(74.8%).

Discussion

From the results, it is found that among the three methods: support vector machine, Logistic Regression, support vector machine with Logistic Regression when used support vector machine with Logistic Regression perform well in all aspects.

In contrast with other previous work on drug testing obtained from the review

website (druglib.com) this approach is showing better accuracy. Felix Graber et al in 2018 worked aspect-based sentiment analysis of product ratings with cross-domain and cross-data learning by used support vector machine algorithm and they got accuracy (70.6%) by using same dataset in this work. When comparing between two works, find that improving classification using support vector machine and logistic regression gave higher accuracy results than previous work that relied on only support vector machine.

Conclusion

With the increasing growth of user-generated content on the Internet, the study of sentiments in digital libraries is becoming important. Sentiment analysis was applied to the health and medical fields, concentrating in particular on public opinion on drugs with specific aspects

This paper presents improving classification in drugs reviews by using machine learning algorithms. Support vector machine was a highly robust in nature tested with the different parameters of consistency. In general, statistical based approach perform better than Support vector machine in this work. Logistic regression works better and provides higher prediction accuracy. This experimental analysis shows that Logistic Regression could be a possible solution for increasing the classification performance.

By general use, SVM & Logistic Regression. They perform similar in every model with SVMs having an edge over Logistic Regression, because SVM can have a better classification algorithm as it does margin classification while logistic regression classifies based on a level of chance that is not a better classification based on the significant characteristics. Therefore, the two algorithms were linked together in this work to obtain more accurate results.

References

1. Cardie, Claire. "Sentiment Analysis and Opinion Mining Bing Liu (University of Illinois at Chicago) Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, 5 (1)), 2012, 167 pp; paperbound, ISBN 978-1-60845-884-4." 2014, pp511-513.
2. Denecke, Kerstin, and Yihan Deng. "Sentiment analysis in medical settings: New opportunities and challenges." Artificial intelligence in medicine, 2015, Vol.64, No.1, pp17-27.
3. Na, Jin-Cheon, et al. "Sentiment classification of drug reviews using a rule-based linguistic approach." International conference on asian digital libraries. Springer, Berlin, Heidelberg, 2012.
4. Na, Jin-Cheon, and Wai Yan Min Kyaing. "Sentiment analysis of user-generated content on drug review websites." Journal of Information Science Theory and Practice 3.1 2015, pp6-23
5. Alhuzali, Hassan, and Sophia Ananiadou. "Improving classification of adverse drug reactions through using sentiment analysis and transfer learning." Proceedings of the 18th BioNLP Workshop and Shared Task. 2019.
6. Alatabi, Hayder A., and Ayad R. Abbas. "Sentiment Analysis in Social Media using Machine Learning Techniques." Iraqi Journal of Science 2020, pp193-201.
7. Abd, Dhafar Hamed, Ayad R. Abbas, and Ahmed T. Sadiq. "Analyzing sentiment system to specify polarity by lexicon-based." Bulletin of Electrical Engineering and Informatics, 2010, Vol.10 No.1,pp 283-289.

8. Goeuriot, Lorraine, et al. "Sentiment lexicons for health-related opinion mining." Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, 2012.
9. Leaman, Robert, et al. "Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts in health-related social networks." Proceedings of the 2010 workshop on biomedical natural language processing, 2010.
10. Yazdavar, Amir Hossein, Monireh Ebrahimi, and Naomie Salim. "Fuzzy based implicit sentiment analysis on quantitative sentences." arXiv preprint arXiv: 1701.00798, 2017.
11. Gopalakrishnan, Vinodhini, and Chandrasekaran Ramaswamy. "Patient opinion mining to analyze drugs satisfaction using supervised learning." Journal of applied research and technology, 2017, Vol15, No.4 pp311-319.
12. Gräßer, Felix, et al. "Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning." Proceedings of the 2018 International Conference on Digital Health, 2018.
13. Zhang, Min, and Guohua Geng. "Adverse Drug Event Detection Using a Weakly Supervised Convolutional Neural Network and Recurrent Neural Network Model." Information, 2019, Vol.10, No.9.
14. Vijayaraghavan, Sairamvinay, and Debraj Basu. "Sentiment Analysis in Drug Reviews using Supervised Machine Learning Algorithms." arXiv preprint arXiv, 2020.
15. Biseda, Brent, and Katie Mo. "Enhancing Pharmacovigilance with Drug Reviews and Social Media." arXiv preprint arXiv, 2020.
16. Cristianini, Nello, and John Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.
17. Schölkopf, Bernhard, Alexander J. Smola, and Francis Bach. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2002.
18. Dreiseitl, Stephan, and Lucila Ohno-Machado. "Logistic regression and artificial neural network classification models: a methodology review." Journal of biomedical informatics 2002, Vol.35.No.5, pp352-359.
19. Chaudhuri, Kamalika, and Claire Monteleoni. "Privacy-preserving logistic regression." Advances in neural information processing systems 2009.
20. Rymarczyk, Tomasz, et al. "Logistic Regression for Machine Learning in Process Tomography." Sensors 2019.