Opinion Mining to Analyze Drug Satisfaction of Patients

Dr. C. S. KanimozhiSelvi[1] V. M. Nishok[2] D. Pavithra[3] R. Pavithra[4]
[1,2,3,4]Department of Computer Science & Engineering
[1,2,3,4]Kongu Engineering College, Perundurai, India

Abstract— People express themselves by giving opinions, feedback, suggestions or ideas about any object. Opinions can be expressed in many ways such as it can be expressed on twitter, Facebook, reviews, blogs etc. World and technology is growing the faster rate so for taking decisions such as buying a product, voting for a politician etc. people are using opinion present on blogs, social networking sites or shopping websites. In opinion mining, most of the researchers have worked on general domains such as electronic products, movies, and restaurants reviews but not much on health and medical domains. Patients using drugs are often looking for stories from patients like them on the internet which they cannot always find among their friends and family [1].The opinion mining method employed in this work focuses on predicting the drug satisfaction level among the other patients who already experienced the effect of a drug. Sentiment of each reviews are analysed and classification algorithms Naïve Bayes and k-nearest neighbour are applied to predict the satisfaction level of patients. In this paper we have taken two drugs of same disease. Polarity of each dataset is analysed to find the best one.

Key words: Opinions, Opinion Mining, Sentiment

I. INTRODUCTION

Opinions are statements that reflect people's perception or sentiment. These statements also provides opinion on objects or events. Opinion Mining or Sentiment analysis is a task under natural language processing for finding the mood of the customers about a purchasing of a particular product or topic. It involves building a system to collect and examine opinions about the product made in many online purchasing sites [1][17].

Drug surveillance is a major factor of drug safety once a drug has been released to the public use. Drug trials are often done in limited number of test subjects where the probability to detect uncommon adverse effects is minimal. Also the volunteers or patients participating in drug trials are also different from those receiving licensed medications, differing in age, co-morbidity and poly-pharmacy. Thus it is necessary to study the safety of marketed drugs on an epidemiological scale. It is also important to understand how the general population uses a particular drug, perceive its safety, reactions and efficiency.[2] The objective of this work is to analyse the drug satisfaction level of patients by using supervised learning algorithms. This paper is outlined as follows. Section 2 narrates the related work. Section 3 discusses the methodology used to develop the models. The data source used is reported in Section 4. The various classification methods used are introduced in Section 5. Section 6 discusses about the results. Section 8 concludes the work.

II. RELATED WORK

The automatic analysis of user generated contents such as online news, reviews, blogs and tweets can be extremely valuable for tasks such as mass opinion estimation, corporate reputation measurement, political orientation categorization, stock market prediction, customer preference and public opinion study.

In sentiment classification, a classifier is trained using labelled data, annotated from the domain in which it is applied. Pang et al (2002) examined whether it is sufficient to treat sentiment classification simply as a special case of topic-based categorization or whether special sentiment-categorization methods need to be developed. This approach used three standard algorithms: Naive Bayes classification, maximum entropy classification, and support vector machines (SVMs) for sentiment classification. In topic-based classification, all three classifiers have been reported to achieve accuracies of 90% and above for particular categories. This shows that sentiment categorization is more difficult than topic classification.[9]

Turney (2002) measured the co-occurrences between a word and a set of manually selected positive words (e.g., good, nice, excellent and so on) and negative words (e.g., bad, nasty, poor and so on) using pointwise mutual information to compute the sentiment of a word [13].

Liu et al (2004) proposed a method to summarize all the customer reviews of a product. This summarization task is different from traditional text summarization because the users are much interested in the specific features of the product that customers have opinions on and also whether the opinions are positive or negative. Hence, the approach does not summarize the reviews by selecting or rewriting a subset of the original sentences from the reviews to capture their main points as in the classic text summarization. It focuses on mining opinion/product features that the reviewers have commented on. The drawback is that there is no group features according to the strength of the opinions that have been expressed on them [14].

Kanayama et al (2006) proposed an approach to build a domain-oriented sentiment lexicon to identify the words that express a particular sentiment in a given domain. By construction, a domain specific lexicon considers sentiment orientation of words in a particular domain. Therefore, this method cannot be readily applied to classify sentiment in a different domain.

Ding et al (2008) focused on customer reviews of products. In particular, the author reviewed the problem of determining the semantic orientations (positive, negative or neutral) of opinions expressed on product features in reviews. So, the author proposed holistic approach that can accurately infer the semantic orientation of an opinion word based on the review context. It provided a new function which is used to combine multiple opinion words in the same sentence. [4]

Pang et al (2008) focused on the methods that seek to address the new challenges raised by sentiment aware applications, as compared to those that are already present in more traditional fact based analysis. This paper includes a material on summarization of evaluative text and on broader

issues regarding privacy, manipulation, and economic impact that the development of opinion oriented information access services gives rise to. To facilitate future work, a discussion of benchmark datasets is also provided. [9]

Ramage et al (2009) introduced Labeled LDA, a topic model that constraints Latent Dirichlet Allocation by defining a one–to–one correspondence between LDA's latent topics and user tags. This allows Labeled LDA to directly learn word tag correspondences. This work demonstrates Labeled LDA's improved expressiveness over traditional LDA with visualizations of a corpus of tagged web pages from del.icio.us. Labeled LDA outperforms SVMs by more than 3 to 1 when extracting tag specific document snippets. [12]

Zhang et al (2010) focused on mining features. Double propagation works well for medium-size corpora. However, for large and small corpora, it can result in low precision and low recall. To deal with these two problems, two improvements based on part-whole and "no" patterns are introduced to increase the recall. Then feature ranking is applied to the extracted feature candidates to improve the precision of the top-ranked candidates. It can rank feature candidates by feature importance which is determined by two factors: feature relevance and feature frequency. [14]

Daume et al (2010) proposed a semi-supervised (labeled data in source, and both labeled and unlabeled data in target) extension to a well-known supervised domain adaptation approach. This semi-supervised approach to domain adaptation is extremely simple to implement, and can be applied as a pre-processing step to any supervised learner. However, despite their simplicity and empirical success, it is not theoretically apparent why these algorithms perform so well. Compared to single-domain sentiment classification, cross-domain sentiment classification has recently received attention with the advancement in the field of domain adaptation. [3]

Raimon et al (2013) dealt with analyzing short messages about brands in twitter trying to classify them between positive and negative using Sentiwordnet. After several experiments, a semi supervised approach is applied and the quality of the dictionary is improved to adapt it to a specific domain. Also, the relevant content inside those tweets are analyzed to know the reason why something is positive or negative. Due to the lack of strong grammatical structures inside tweets, an approach based on structured N–grams is proposed. For that, a new idea called sentigram is modeled that consists of the aggregation of several N–grams. This approach allows creating models very precise to specific domains and at the same time capturing the relation between aspects and sentiment words.

Quan et al (2014) identified product features from reviews as well as a bottleneck in feature-level sentiment analysis. This study proposed a method of unsupervised product feature extraction for feature-oriented opinion determination. The domain-specific features are extracted by measuring the similarity distance of domain vectors. A novel term similarity measure (PMI–TFIDF) is introduced to evaluate the association of candidate features domain entities.[11] of candidate features domain entities.[11]

III. METHODOLOGY

The following is the summary of our methodology for developing and validating the prediction models. The work flow diagram is shown in figure 1.[2]

1) Perform data pre-processing
2) Construct the document term matrix for the drug reviews selected for drug I and drug II.
3) Apply the following classification methods using the respective training data set
   a) Naive Bayes
   b) k-nearest neighbour
4) Predict the class (positive or negative) of each review in the test data set.
5) Compare the prediction results with actual values of candidate features domain entities.[11]

IV. METHODOLOGY

The following is the summary of our methodology for developing and validating the prediction models. The work flow diagram is shown in figure 1.[2]

1) Perform data pre-processing
2) Construct the document term matrix for the drug reviews selected for drug I and drug II.
3) Apply the following classification methods using the respective training data set
   a) Naive Bayes
   b) k-nearest neighbour
4) Predict the class (positive or negative) of each review in the test data set.
5) Compare the prediction results with actual values.

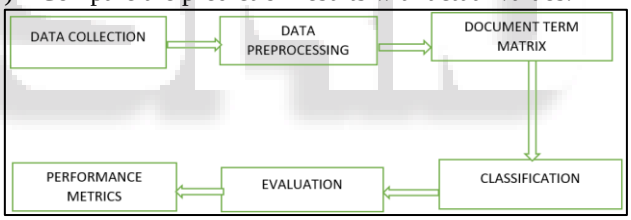

Fig. 1: Work Flow for Classification

For analysing the sentiment of reviews the steps are used as follows:

1) Perform data pre-processing
2) Create word cloud for pre-processed data with minimum frequency
3) Use syuzhet package for analysing the sentiments
4) Plot the sentiments and its values

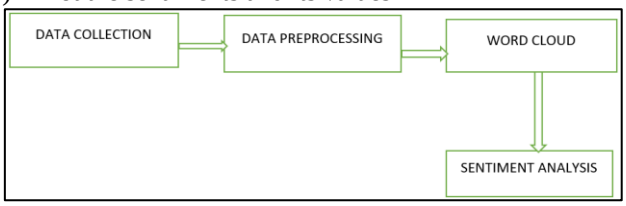

Fig. 2: Flow Diagram for Sentiment Analysis

The above diagram shows the workflow for analysing the sentiments of reviews.

V. DATA SOURCE

We constructed the drug review dataset from the popular web resource www.askapatient.com. Figure 3 shows the sample review format from www.askapatient.com. Opinion mining

is proved to be domain specific due to the nature of the language used. But the focus of this work is on opinion mining in singled domain i.e. drug reviews. We constructed two datasets (Dataset I and Dataset II) by collecting the reviews of two popular drugs such as Cymbalta and lamtical. Cymbalta is used to treat major depressive disorder in adults. Lamtical is also used to treat depressive disorders. These two drugs differ in their nature of use. Each user comment is stored as separate text document, which are considered as samples for training and testing in learning process. [2]

Rating	Reason	Side effects	comments	Sex	Age	Duration	Date
4	Depression fibromyalgia	Heartarrhythmia (skippingbeats) Insomnia	Was amazed & Delighted how quickly symptom disappeared in the first week. I have struggled to lose since menopause	F	59	6 weeks30mg	5/23/15

Fig. 3: Sample Dataset

VI. ALGORITHMS USED

A. *Naive Bayes*

Naive Bayes is a simple technique for constructing classifiers, models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle. All naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.[15]

1) BAYES Theorem

$P[H\ E] = P[E\ /H] \times P[H] / P[E]$

- H – hypothesis
- E-evidence related to the hypothesis H, i.e., the data to be used for validating (accepting/rejecting) the hypothesis H
- P(H) – probability of the hypothesis (prior probability)
- P(E) – probability of the evidence i.e., the state of the world described by the gathered data
- P(E|H) – (conditional) probability of evidence E given that the hypothesis H holds
- P(H|E) – (conditional) probability of the hypothesis H given the evidence

After calculating the posterior probability for a number of different hypotheses, we can select the hypothesis with the highest probability. This is the maximum probable hypothesis and may formally be called the maximum a posteriori (MAP) hypothesis.

This can be written as:

$$MAP(h) = \max(P(h|d))$$

The P(d) is a normalizing term which allows us to calculate the probability.

B. *K-Nearest Neighbour*

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its *k* nearest neighbours (*k* is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor. *k*-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is among the simplest of all machine learning algorithms. The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the classification phase, k is a user-defined constant, and an unlabelled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point. [16][18]

VII. RESULTS & DISCUSSION

In general, the process of prediction contains four different results called true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The confusion matrix displays these four results. Column A presents the tested positive results, and column B presents the tested negative results. The first row shows the predicted results for the positive class, and the second row shows the predicted results for the negative class.

From the outcome of detailed accuracy, we present some significant indicators as follows. The precision is calculated by

$$\text{Precision} = \frac{TP}{TP+FP}$$

The recall, also known as the specificity, is calculated by

$$\text{Recall} = \frac{TP}{TP+FN}$$

After the application of Naïve Bayes and K nearest neighbour (knn), it is seen that the algorithm KNN gives the best accuracy for both the datasets. Thus it shows that KNN is the best algorithm which gives the best result for both Cymbalta and Lamtical drug datasets.

The plot for sentiment analyse of Cymbalta dataset is shown in the figure iv.

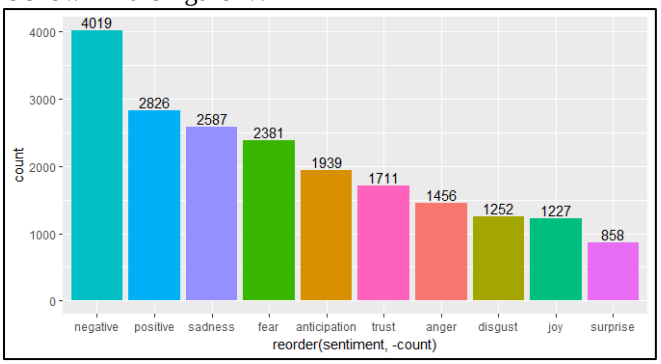

Fig. 4: Sentiment Analysis of Dataset 1

The sentimental analysis for Lamtical dataset is as follows:

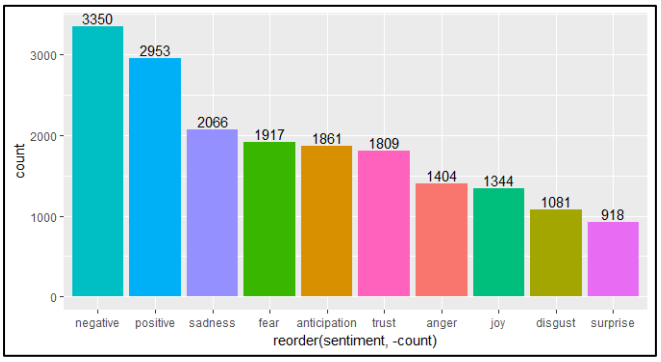

Fig. 5: Sentiment Analyse for Dataset 2

From figure 4 and 5 it is clear that both datasets have high negative polarity than positive polarity. Comparing Dataset 1 and Dataset 2 from the plots we can say that dataset 2 ie.)Lamtical is more popular among customers than Cymbalta ie) dataset 2. Because Cymbalta has more negative polarity compared to Lamtical.

VIII. Conclusion

Opinion mining and sentimental analysis is an emerging field of data mining used to extract the knowledge from a huge volume of customer comments, feedback and reviews on any product or topic etc. A lot of work in opinion mining in customer reviews has been conducted to mine opinions in form of document, sentence and feature level sentiment analysis. Thus in our work we have compared two drug datasets of same disease to find the best one by comparing the polarity of them. And also the classification techniques Naïve Bayes and K nearest neighbor are applied to predict the satisfaction level of patients. In future, Opinion Mining can be carried out on set of discovered feature expressions extracted from reviews. The Opinion Mining and in natural language processing community, Sentiment Analysis become a most interesting research area. A more innovative and effective techniques needed to be invented which should overcome the current challenges faced by Opinion Mining and Sentiment Analysis.[1][17]

References

[1] Opinion Mining and Sentiment Analysis on Customer Review Documents- A Survey Surya Prakash Sharma1, Dr Rajdev Tiwari2, Dr Rajesh Prasad3

[2] www.jart.ccadet.unam.mx,Patient opinion mining to analyze drugs satisfactionusing supervised learningVinodhini Gopalakrishnan∗, Chandrasekaran Ramaswamy

[3] Daumé III, Hal, Abhishek Kumar and Avishek Saha (2010), 'Frustratingly easy semi-supervised domain adaptation', In Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing, Association for Computational Linguistics, pp.53-59.

[4] Ding, Xiaowen, Bing Liu and Philip S. Yu (2008), 'A holistic lexicon-based approach to opinion mining', In Proceedings of the 2008 International Conference on Web Search and Data Mining, Association for Computing Machinery, pp. 231-240.

[5] Hu, Minqing and Bing Liu (2004), 'Mining opinion features in customer reviews', In Proceedings of the national conference on artificial intelligence, Vol.4, No.4, pp.755-760.

[6] Hu, Minqing and Bing Liu (2004), 'Mining and summarizing customer reviews', In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp.168-177.

[7] Kanayama, Hiroshi and Tetsuya Nasukawa (2006), 'Fully automatic lexicon expansion for domain-oriented sentiment analysis', In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp.1-9.

[8] Marrese-Taylor, Edison, Juan D. Velásquez and Felipe Bravo-Marquez (2014), 'A novel deterministic approach for aspect-based opinion mining in tourism products reviews', Expert Systems with Applications, Vol.41, No.17, pp.7764-7775.

[9] Pang, Bo, Lillian Lee and Shivakumar Vaithyanathan (2002), 'Thumbs up?: sentiment classification using machine learning techniques', In Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Vol.10, pp. 79-86.

[10] Pang, Bo and Lillian Lee (2008), 'Opinion Mining and Sentiment Analysis', Foundations and Trends in Information Retrieval, Vol. 2, No. 1/2 pp. 1-135.

[11] Quan, Changqin and Fuji Ren (2014), 'Unsupervised product feature extraction for feature-oriented opinion determination', Information Sciences, Elsevier, pp.16-28.

[12] Ramage, Daniel, David Hall, Ramesh Nallapati and Christopher D. Manning (2009), 'Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora', In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics,Vol.1, pp.248-256.

[13] Turney and Peter D (2002), 'Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews', In Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417-424.

[14] Zhang, Lei, Bing Liu, Suk Hwan Lim and Eamonn O'Brien-Strain (2010), 'Extracting and ranking product features in opinion documents', In Proceedings of the 23rd international conference on computational linguistics: Posters, Association for Computational Linguistics, pp. 1462-1470.

[15] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[16] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

[17] Kalaichelvi C, Selvi K. Frequent item sets generation using collective support threshold for associative classification. In National Conference on Recent Trends in Communication and Signal Processing 2009 (Vol. 2009).

[18] Kanimozhiselvi CS, Tamilarasi A. Mining of High Confidence Rare Association Rules with Automated Support Thresholds. European Journal of Scientific Research. 2011.

