

Opinion Mining for Drug Reviews

J.Satish Babu, Chinnam Siva Koteswara Rao, Debrup Banerjee, S. Sagar Imambi,
G Krishna Mohan

Abstract: Nowadays, people extract information from many user-centered platforms like Facebook, Twitter, and Amazon etc. Even the information about the medicines to be used is being extracted from the net rather than getting appointment from a doctor which takes a lot of time. Probabilistic aspect for mining models (PAMM) is a technique which is used to identify the medicines which belong to the same class rather than searching the medicines from all classes. Another technique used in this is Opinion Mining which extracts only specified information from large amounts of data. Compared with various products, medicines have limited aspects like reviews from the people, their experiences, type of medicine to be used, medicine dosage etc. PAMM technique works more efficiently compared to other techniques.

Index Terms: User Centered Platforms, Probability, Opinion Mining and Aspect Based Mining.

I. INTRODUCTION

With the wide increase in the usage of Internet, people are depending upon Internet even for purchasing different types of products. They use many user centered platforms like Facebook, Twitter, and Amazon etc. These user centered platforms are used by a massive amount of people. They are very smart that they purchase a particular product by comparing it with another product of the similar kind or by viewing comments or reviews given by various users on that particular product. As the customers and the usage of Internet increases, it is difficult to use concepts like text mining.

Nowadays, people are even purchasing medical products online instead of waiting for a long time in the hospitals to get appointment for consulting doctors. They are mainly depending upon the comments or reviews given by other users who have used that particular medicine earlier or who are suffering from the same problem. But the problem here is, the same medicine may not give side effects for men but it may give side effects for women.

Opinion Mining is a technique which extracts information based on the comments given by various users. It extracts useful information from large amounts of information provided. Automatic extraction of customers' opinion can benefit both manufacturers and customers. This type of

mining can be done for supervised as well as unsupervised approaches.

Previous studies deal about products like electronic gadgets, household items, some aspects of medicines like ease of use, side effects, dosage etc. They use aspect based mining where some part or a particular word is taken as a reference from the whole sentence. Due to this, the difference between the side effects caused by different age groups has become a challenge. This is solved using opinion mining where the whole sentence is taken as a reference.

Merchants promoting the products on the online shopping stores are increasing very rapidly and the main motto regarding the usage of web based is to increase the customers globally. As Online purchasing are becoming a daily routine to the customers for the products and becoming trendy and the vendors are very keen in knowing the interest of the customers by their purchase pattern and reviews given by them. Sometimes it creates a lot of ambiguity to the customers and merchants to track the products of their choice and how much production is required to cope up the demand of the customers and customers fail to choose their product of their preferred brand. To overcome this, a better learning system which exclusively focuses on the choices of the customer to be known for the better decision making of the Merchants by the customer reviews.

Our task is performed in three Steps.

- 1 Products of the customer choice to be mined and listed.
- 2 Distinctive Opinion based on the review of the customers to make a decision on the product.
- 3 Consolidate and store the Results.

With the flourish of the net, on-line review is changing into an additional and additional helpful and necessary data bank for individuals. As a result, after performing the mining on the review data a better analysis gives a better result. It helps in deciding whether or not the opinions square measure positive or negative. This work mainly focuses on the domain related to the drugs which helps the patients for better selection of their respective drug.

II. LITERATURE REVIEW

Probabilistic aspect mining model is used to review drugs based on the side effects, ease of use, dosage by calculating conditional probability, entropy of various drugs [1]. Opinions are extracted from various data sources such as blogs etc. To implement opinion mining on these sources many processes are implemented [2]. Evolutionary computation has been utilized in usual process, starting from efforts to induce grammars to models

Revised Manuscript Received on May 07, 2019.

J.Satish Babu, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India.

Chinnam Siva Koteswara Rao, Department of Information Technology, VFSTR deemed to be University, Guntur, Andhra Pradesh, India.

Debrup Banerjee, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India.

S. Sagar Imambi, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India.

G Krishna Mohan, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India.

of language development through parameter optimisation and search [3]. Probabilistic feature mining model is use for distinctive the aspects/topics regarding category labels. This reduces the prospect of getting aspects created from the compounding ideas of various categories [4].Opinion Mining is widely used area to analyze the end-user opinion based on their preferences, likes, emotions and also considering their sentiments. It is the most appropriate area for forecasting data and predicting the need of the end-users. and is additionally wide studied in data processing, internet mining and text mining [5].The various techniques of opinion mining like Naïve Thomas Bayes theorem, Support Vector-Machine(SVM),Multilayer Perception,cluster square measure enforced[6].

III. THEORETICAL ANALYSIS

The information gathered from the massive amount of reviews is collected. Various models such as Probabilistic aspect mining model (PAMM) can be applied. Probabilistic aspect mining model is used to review drugs based on the side effects, ease of use, and dosage by calculating conditional probability, entropy of various drugs. These approaches, unfortunately, agonize from a extreme problem, it is difficult to hold close the underlying aspects or principles from simply a set of words correlated with a type label. There is no intuitive algorithm to crew the words so that each crew conveys one or a few without difficulty understandable concepts. PAMM model is efficient when compared to NFM, LDA, sLDA, SSMF and DiscLDA. But it does not differentiate between the side effects caused by different age groups and different gender.

As several drug evaluate web sites Square measure operational with ranking functions, prediction of sentiments instead of representing for clustering of the elements based on their categories. As an example, if the evaluation Square measure related gender data, individuals should also be inquisitive about knowledge the elevation big difference between feminine patients and male patients.

This venture is completely unique from established aspect-based opinion mining inside which the undertaking targets to extract all elements and their sentiments from reviews. Bearing on the subject definition, no longer all the factors on the other hand solely relevant elements have to be compelled to be extracted. Sometimes, an aspect would possibly have to be divided extra (in finer granularity) as a result of totally restrained factors of its square measure needed. As an example, considering the side of facet consequences of a drug, male suffers are additionally anxious a couple of unique side result where as alternative side impacts Square measure of much less involved.

Opinions are necessary as a result of whenever we'd like to create a choice, we would like to listen to others' opinions. Users' opinions play a significant role. Opinions may be conveyed by adjectives, nouns, verbs and adverbs. They're characterized by polarity (+,-) and strength (great>good).Opinion mining may be a subjective analysis that deals with have an effect on, opinions, sentiments.

IV. EXPERIMENTAL INVESTIGATIONS FOR OPINION MINING ALGORITHMS

A. Naïve Bayes theorem

If there are 2 activities say, e1 and e2 then the chance of prevalence of match e1 when e2 has already occurred is given by way of the subsequent mathematical formula:

$$P(e1|e2)=(P(e2|e1)P(e1))/e2$$

This algorithmic program is enforced to calculate the chance of information to be positive or negative. So, Conditional likelihood of a sentiment is given as:

$$P(\text{Sentiment}|\text{Sentence})=((P(\text{Sentiment})P(\text{Sentence}|\text{Sentiment}))/P(\text{Sentence}))$$

Conditional chance is given as:

$$P(\text{Word Sentiment})=(\text{Number of words passed off} +1)/(\text{number of phrases belonging to that class total number of words})$$

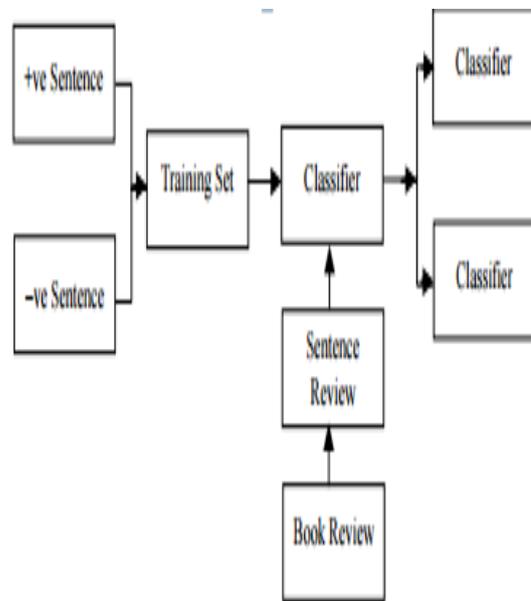


Fig. 1 Algorithm for Naïve Bayes.

Evaluating the algorithm:

The following measures are required to evaluate the algorithm:

- Accuracy
- Precision
- Recall
- Relevance

The contingency table for the algorithm is as follows:

Table . 1 Contingency table

	Relevant	Irrelevant
Detected Opinions	True Positive (tp)	False Positive (fp)
Undetected Opinions	False Negative (fn)	True Negative (tn)

Precision= $tp/(tp+fp)$

Accuracy= $(tp+tn)/(tp+tn+fp+fn)$

Recall= $tp/(tp+fn)$

Advantages of Naïve Bayes algorithm:

- It is very easy to implement.
- The model has efficient computation.
- Disadvantages of Naïve Bayes theorem:
- Assumption of attribute being freelance, which cannot be there essentially applicable

A. Support Vector Machine (SVM)

SVM ought to be a supervised learning model. This model is expounded to a learning algorithm that analyzes the statistics and identifies the pattern for classification. The construct of SVM formula is based totally on choice plane that defines boundaries .A selection plane separated crew of instances having special classification memberships.

$$D=\{(x_i,c_i)|x_i \in R^p, c_i \in \{-1,1\}\} n_i-1$$

Where, X_i may be a p-dimensional actual vector perceive the most -margin hyper craft i.e. splits the factors having $c_i = 1$ from these having $c_i = -1$. Any hyperplane may also be written as a result of the set of things satisfying

$$w \cdot x - b = 1$$

B. Extensions of SVM

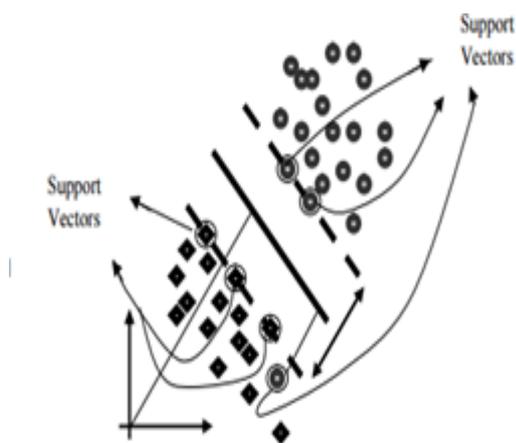


Fig. 2 Principle of SVM

C. Multiclass SVM

Basically SVM relevant for two class tasks aside from the multiclass problems there is multiclass SVM is procurable. at intervals the multi-class case label area unit envisages to things that area unit has drawn beginning a restricted position of assorted parts.

Advantages of SVM

- It offers superb performance.

Disadvantages of SVM:

- Interpretation is troublesome.
- It ought to be pre-processed if missing values occur.

D. Multi-Layer Perceptron (MLP)

Multi-Layer perceptron might be a feed-forward neural network, with one or N layers amongst inputs and output. Feed forward suggests that i.e. uni-direction drift of statistics like from enter layer to output layer. This ANN that multilayer perceptron start with input layer anywhere each node suggests that a grammatical construction variable. Input nodes or vegetative cells unit coupled with each vegetative cell in future layer.

The output layer is created up as follows:

When the prediction is binary output layer created of 1 physical telephone and when the prediction is non-binary then output layer created from N Vegetative cell.

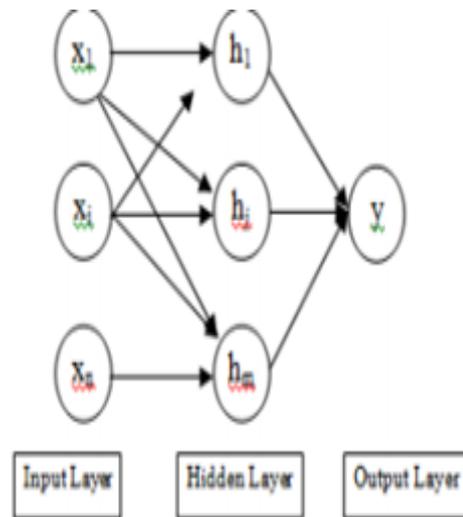


Fig. 3 Structure of MLP

MLP has two phases:

Phase1:

It is the forward half where activation is propagated from the input layer to the output layer.

Phase2:

In this, to vary burden and bias extremely price errors amongst wise & actual prices and also the requested value within the output layer is propagate at intervals the backward direction.

MLP is a normal technique owing to the particular reality i.e. it'll act as universal perform approximator. MLP may be a general, versatile and non-linear tool as a result of a “back propagation” network has minimum one hidden layer with several non-linear entities which is able to study each operates or relationship between a team of entering



and output variable (whether variables region unit is distinct or continuous). An advantage of MLP, compare to classical modeling technique is that it does not enforce any kind of constraint with relevancy the initial data neither can it usually begin from specific assumptions.

Another advantage of the plan of action lies in its capability to analysis good models even notwithstanding the presence of noise at intervals the analyzed information, as arises once there is also associate existence of neglected and outlier values at intervals the spreading of the variables. Hence, it is a durable approach once addressing issues of noise at intervals the given info.

Advantages of MLP

- MLP can analyze each and every and each relationship between input and output variables.

Disadvantages of MLP

- MLP desires longer for execution examine to the specific technique as a finished result of flexibility lies at intervals the requirement to possess sufficient teaching data.

E. Stages of Opinion mining

Data assortment and pre-processing:

In this stage, it's nonmonogenic the text which will be analyzed for detection of opinions. it's necessary, per the methodology used, to get rid of all matters that not categorical opinions. Throughout this half, pre-processing is finished to eliminate supernumerary words or unsuitable opinions. It's necessary to extract keywords from the text which could provide associate correct classification. These keywords unit sometimes hold on as associate array of choices $A = (A1, A2, \dots, An)$. each a part of array might be a word from the primary text, called side feature. for every feature, can exist a binary worth, indicating the presence or absence of the feature, or a worth that expresses the frequency of look at intervals the text. Selecting common aspects is crucial so they categorical those relevant opinions for sentiment analysis.

F. Classification

In this half, content polarity is understood. Sometimes three classes' area unit used for classification: positive, negative and neutral. Classification algorithms used for sentiment analysis, reckoning on the plan of action used supervised or unsupervised, need coaching set with pre-marked examples. It's essential to educate the model used for classification with domain-specific data. Marking is completed by expressing judgment and polarity of employment sets.

G. Aggregation

At this stage, the opinions classification technique result obtained at the previous stage is subjected to a way of aggregation once some algorithms to ascertain the ultimate opinion of the analyzed text. Presentation is going to be done directly expressing sentiment matter or exploitation graphics.

V. EXPERIMENTAL RESULTS

This section gives the experimental results obtained by various opinion mining algorithms. The opinion mining algorithms are applied on a single drug and the experiment results are as follows:

Table. 2 Experiment result of Naïve Bayes

	Naïve Bayes Multinomial	Naïve Bayes Multinomial Text
Initial %	48	54
Stopwords	53	56
Stemmer	54	59
Unigrams	59	61
Min word frequency from 5-10	65	66
Bigrams	70	71
Trigrams	63	62
Business and user id	63.5	63
Business id	74	75
User id	74	74.7
Attribute selection filter	78	78.4
Bag of words count	79	
Overall Accuracy	79.49	79.6

The Naïve Bayes theorem is easier to implement and while execution it performs an efficient computation compared to other algorithms. But as the assumptions of attributes are independent, sometimes they are not valid.

The model Support Vector Machine came into existence and the main advantages of this algorithm is it has good performance and highly independent on the data set dimensionality.

Table. 3 Experiment Results of SVM

M%	100%		70%		50%		30%		10%				
	LS	E1 E2 LS	E1 E2 LS	E1 E2 LS	E1 E2 LS	E1 E2 LS	E1 E2 LS	E1 E2 LS					
A+B	822.3	850.4	830.6	836.3	838.2	832.4	832.7	857.0	843.5	849.9	971.9	957.0	938.9
A+B+C	835.3	848.0	837.6	838.3	893.5	833.6	833.5	856.7	848.1	848.9	982.1	957.2	942.3
A+B+C+D	846.5	867.1	851.8	855.8	834.6	876.3	853.5	951.4	876.7	859.9	980.6	986.8	960.3
A+B+D	853	878.2	853.4	852.8	890.1	853.3	851.5	904.0	853.1	868.1	973.4	953.5	945.6
A+B+E	853.8	837.5	860.2	851.6	882.2	855.9	852.0	880.5	872.1	868.1	995.8	964.8	940.7
A+B+D+E	868.1	880.6	889.3	885.3	882.1	909.5	893.1	860.7	895.8	890.4	979.0	991.9	963.6
A+B+C+D+E	903.6	882.7	879.1	875.2	867.8	915.1	884.9	855.5	891.3	885.7	984.2	991.9	963.6
A+B+C+E	849.1	865.6	851.9	855.8	854.0	876.3	853.5	926.4	876.7	859.9	975.9	986.8	960.3

But using SVM may lead to preprocessing in case of missing values and it is very difficult to implement.

VI. CONCLUSION

Using opinion mining many drug reviews are analyzed easily. By using the mining techniques the discrimination between the side effects caused by different age groups and genders can be identified. This also differentiates the side effects caused by different people who have different problems.



For future use, Opinion Mining should control the usage of fake reviews which are very common these days. The tools of opinion mining should be used efficiently so that reviews from various sites, blogs are collected and together analyzed. For future work, opinions are to be collected from different sites and the problems of opinion mining are to be reduced like reduction of fake reviews, eliminating repeated reviews etc.

REFERENCES

1. Cheng, Victor C., Clement HC Leung, Jiming Liu, Alfredo Milani, "Probabilistic aspect mining model for drug reviews", IEEE transactions on knowledge and data engineering, Vol. 26, No. 8, 2014, pp.2002-2013.
2. Angulakshmi, G., R. ManickaChezian, "An analysis on opinion mining: techniques and tools", International Journal of advanced research in computer and communication engineering, Vol. 3, No. 7 2014, pp. 2278-1021.
3. Bungum, Lars, Bjorn Gamback, "Evolutionary algorithms in natural language processing", Proceedings of the Norwegian Artificial Intelligence Symposium, pp. 7-19. 2010.
4. Mall, Madhuri, V. A. Chakkarwar, "Aspect Mining Model in Support of Drug Reviews".
5. Dhokrat, Asmita, Sunil Khillare, C. Namrata Mahender, "Review on techniques and tools used for opinion mining", International Journal of Computer Applications Technology and Research, Vol. 4, No. 6, 2015, pp. 419-424.
6. Singh, Pravesh Kumar, Mohd Shahid Husain, "Methodological study of opinion mining and sentiment analysis techniques", International Journal on Soft Computing, Vol. 5, No. 1, 2014, pp. 11.
7. Blair-Goldensohn, Sasha, Kerry Hannan, Ryan McDonald, Tyler Neylon, Jeffrey C. Reynar, "Phrase based snippet generation", U.S. Patent 8,010,539, issued August 30, 2011.
8. Khairnar, Jayashri, Mayura Kinikar, "Machine learning algorithms for opinion mining and sentiment classification", International Journal of Scientific and Research Publications, Vol. 3, No. 6, 2013, pp. 1-6

AUTHORS PROFILE



J. Satish Babu, working as Assistant Professor in the Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, A.P., India. His area of interests are Data Mining, Software Engineering. Email: jampanisatishbabu@kluniversity.in.



Chinnam Siva Koteswara Rao, working as Assistant Professor in the Department of Information Technology, VFSTR deemed to be University, Guntur, Andhra Pradesh, India, His area of interests are Data Mining, Image Processing. areEmail: csrvictory@gmail.com.



Debrup Banerjee is an Assistant Professor at Koneru Lakshmaiah University, India. He is involved in teaching soft computing, machine learning, artificial intelligence and computer databases. He received the BE degree from Shivaji University in India, the MS degree in Computer Science from Hampton University, USA, and a recent Ph.D. degree in Electrical and Computer Engineering from Old Dominion University, USA, in 2017. His current research interest is applied machine learning. He has also worked in software engineering for a renowned multinational company in the past.