# Probabilistic Mining Model for Drugs Classification in Data Mining

Haritha Paidi[1], L. Prasanna Kumar[2]

[1]*Final M.Tech Student,* [2]*Assoc.Professor*
[1, 2,] *Department of Computer Science & Engineering, Dadi Institute of Engineering & Technology, Anakapalle-531002, A.P., India*

***Abstract****:*
*Nowaday's finding chronic diseases and drugs are becoming more important for supporting the patient resource information. Extracting patient information from the text is most challenging and also critical. So for extracting patient information from these substantial bodies of texts we are using so many opinion mining techniques. In this paper we are extracting information from these substantial bodies of texts using one of the mining models of classification approach. The classification technique used is Naïve Bayesian classifier which is used for finding the causes that occur by using over doses of drugs and also to find the type of side effect that will occur. After completion of classification we are grouping the related drugs which are causing the same side effects by using Word Comparator Clustering algorithm. By implementing this application we can improve the efficiency and also provide more classification accuracy***.**

*Keywords: Data mining, classification,Naive Bayesianclassifier, Clusterization.*

## I. INTRODUCTION

Unlike usage of general products or service of different products, drugs have a limited number of different aspects i.e. ease of usage, price, effectiveness, and dosages of each drug and side effects of each drug. The aspects of drugs are molecular and chemical aspects, but they are not mentioned for finding the classification process. A difficulty for finding the type of drugs causes of type of side effect and people experience are very diverse. In particular the occurrence of side effects of drugs is depending on the symptoms for each drug that is applicable to another drug. So that for finding relative drugs for occurrence of side effects we are implementing opinion mining in data mining.

Now a day's data mining is widely used for the purpose of performing association rule mining, classification approach and clustering of data. By performing association rule mining we can find out frequent item set from the dataset. Data collected from dataset is important for making decision by using classification approach. By implementing clusterization approach we can group thesame type of data sets. So that by performing these concepts we can get more frequent items sets, classified data set and also perform the group of related data set into single group.

In this paper we address mainly two concepts the first one is finding type of side effect will occur by using the drugs in body. Another one is grouping all related side of effects of drugs into group. By implementing those concepts we filter the related side effects of drugs into single group. Instead of finding related drugs we are using some aspects of each drug i.e. dosage of each drug can be effected by in body. Most usage of drugs we can get more side effects in a body. In this paper we are using one of classifier approach for finding type of drugs can occur the side effect in the body. By performing classification approach we get type side effect occur in a body. After performing classification approach we can group the all type side effect into group by using clustering approach.

Most existing text clustering algorithms are designed for central execution. They require that clustering is performed on a dedicated node, and are not suitable for deployment over large scale distributed networks. Therefore, specialized algorithms for distributed and P2Pclustering have been developed, such as [6], [7], [8], [9].However, these approaches are either limited to a small number of nodes, or they focus on low dimensional data only. In distributed environment, nodes are represented by Privacy reserving Distributed clustering algorithm proposed by "S. Jha, L. Kruger, and P. McDaniel", here data can be clustered by grouping the similar type of objects and secure transmission through protocols[4].Perturbation Method of string transformation proposed for privacy preserving clustering technique by using geometric techniques [10].

The rest of the paper is organized is as follows. Section 2 is used to specify the related work of the our proposed system. Section 3 is specifying the existing system. Section 4 is to used specify the implementation procedure of proposed system. Section 5 is to specify the experimental result our proposed system. Section 6 is to specify the conclusion of our proposed system. Section 7 is to specify the reference of our proposed system.

## II. Related Work

So many probabilistic mining approaches are available for performing the association rule [12] mining concepts classifies the data.These approaches, unfortunately, suffer from a severe problem: it is difficult to understand the underlying aspects or concepts from just a set of words correlated with a class label. There is no intuitivealgorithms to group the words so that each groupconveys one or a few easily understandable concepts.In the recent years aspect mining is more popular for performing concepts of association rule mining. In the frequency based approach [13] extract highly frequency word of noun which meets the specified the review the drugs. Another concept is relation based approach [14][15] identifies relation between the all the reviews. In the relational based approach mainly contains two concepts. i.e. classification of drug reviews and grouping all related drugs which causes the side effects. In contrast topic modelling will identify the group of same type of elements into single group. So that the main advantage of aspect mining is to identification and group of related data into single group. ]

In the recent years the topic modelling [16] is more probabilistic approach is understanding corpus. With this approach contains many topics which are represented by performing multi nominal distribution of words and using those words we can sort topics with highly probabilistic. Another aspect mining is aspect and sentiment unification model [17] were proposed for extracting aspect and predict their associated sentiments. By implementing these opinion mining model is may not appropriate to address of existing problem may not be related to specified labels and performance depend on the selection of words. Recently the interest are in mining concepts is topic modelling. Bleiand McAuliffe [18] proposed the supervised LDA (sLDA) that can take care of different forms of supervised information during topic inference. Apart from probabilistic algorithms, deterministic methods for topic modeling such as non-negative matrix factorization (NMF) [19] were also proposed. By decomposing the data matrix into two low rank matrices, topics can be identified. Semi-supervised NMF(SSNMF) [20] is an extension proposed recently to incorporate the supervised information into NMF. The topics identified are more closely related to the supervised information

## III.Existing System

Previous studies of opinion mining usually deal with popular consumer products or services such as digital cameras, books, electronic gadgets, etc. Entities of medical domain are of far less concerned. It may be because patients are minority groups on the Internet and they are only concerned with specific illnesses or drugs that they are experiencing. Furthermore, people tend to solicit opinions from medical professionals rather than patients. Unlike general products or services, drugs have a very limited number of kinds of aspects: price, ease of use, dosages, effectiveness, side effects and people's experiences. There are other more technical aspects such as chemical or molecular aspects, but they are almost not mentioned in drug reviews. A difficulty in dealing with drug reviews is that the wording in describing effectiveness, side effects and people's experiences are very diverse. In particular, side effects are drug dependent: a set of side effect symptoms for a drug is very unlikely applicable to another drug. This impedes some opinion mining approaches based on lexicons. More importantly, authors sometimes do not indicate which aspects they are describing, they just give descriptions of symptoms, feelings and comments.

## IV. Proposed System

The main objective of proposed system is to find the type of effect that will occur on taking over dose of drugs. Another concept is to group all related effects of drug into number of clusters. By implementing those concepts we can find out and group related side effects into one group. In this paper we mainly focus upon how to find the type of side effect that will occur on taking over dose of drugs by using classification approach. The classification approach followed is Naïve Bayesian classifier. After performing classification approach group all related side effect into groups by using Word Comparator Clustering algorithm. The implementation procedure of each algorithm is as follows.

**Classification approach:**

In this module we are finding the type of side effect that will occur by taking over dose of drugs. Before performing the classification approach we are taking training data set for each drug. The training data set will contain information related side effects occurred by taking over dose of drugs. In this paper we consider four types of drugs information and occurrence of side effect by taking over dose of those drugs. So by taking those values we can classify the type of side effects that will occur by taking over dose of drugs. For the Classification mechanism, we are using Naïve Bayesian classifier for classifying the testing dataset with newly formed optimal feature set based Dataset for diabetes. This approach works with corresponding posterior probability of the individual features with respect to the original dataset.

For the classification process we are using Naïve Bayesian classifier for analysing the testing data with the training information. Naïve Bayesian classifier is defined by a set C of classes and a set A of attributes. A generic class belonging to C is denoted by $c_j$and a generic attribute belonging to A as $A_i$. Consider a database D with a set of attribute values and the class label of the case. The training of the Naïve Bayesian Classifier consists of the estimation of the conditional probability distribution of each attribute, given the class.

It is the probability that X is round and red given that we know that it is true that X is an apple

Here P(X) is prior probability =

P (data sample from our set of fruits is red and round)

P(X), P(H), and P(X/H) may be estimated from given data .Use of Bayes Theorem in Naïve Bayesian Classifier

1. Each data sample is of the type

$X= (x_i)$ i =1(1) n, where $x_i$ is the values of X for attribute $A_i$

2. Suppose there are m classes $C_i$,i=1(1)m.

$X \in C_i$iff

$P(C_i|X) > P(C_j|X)$ for $1 \leq j \leq m, j \neq i$

i.e BC assigns X to class $C_i$having highest posterior probability conditioned on X .The class for which $P(C_i|X)$ is maximized is called the maximum posterior hypothesis. From Bayes Theorem

3. P(X) is constant.

- If class prior probabilities are not known, then assume all classes to be equally likely.
- Otherwise maximize.

$P(C_i) = S_i/S$

Problem: computing $P(X|C_i)$ is unfeasible!

(find out how you would find it and why it is infeasible)

4. Naïve assumption: attribute independence

$= P(x_1,\ldots,x_n|C) = \Pi P(x_k|C)$

5. In order to classify an unknown sample X, evaluate for each class $C_i$. Sample X is assigned to the class $C_i$iff

$P(X|C_i)P(C_i) > P(X|C_j) P(C_j)$ *for* $1 \leq j \leq m, j \neq i$ .

After completion of classification approach we are grouping same type of side effects into one group by using Word Comparator Clustering algorithm

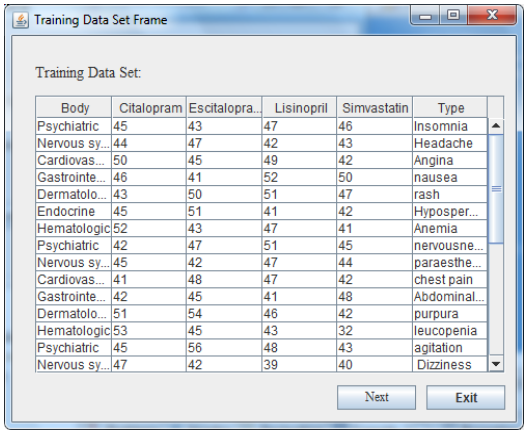
**Word Comparator Clustering algorithm:**

The main objective of this algorithm is to group the same type of side effects into single group. By performing the clusterization we can identify the same type of side effects and the drug which causes these side effects. The implementation procedure of this algorithm is as follows.

1. Specify how many clusters we want. Based on the number of clusters perform the grouping.
2. Randomly choose cluster member based on the number of clusters.
3. Find the equal range of attributes of dataset in set to cluster members.
4. If range of attributes is equal to range of cluster member then that dataset should be kept into that cluster member.
5. Repeat this process until we group the total data set into number of group of clusters.

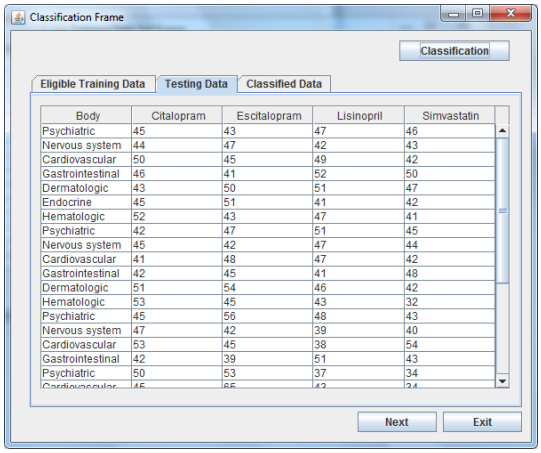
## IV .Experimental Analysis

Our implementation purpose we have used language java and some synthetic datasets for analysis, the following representation shows the complete implementation of the architecture.




| Body | Citalopram | Escitalopra... | Lisinopril | Simvastatin | Type |
|---|---|---|---|---|---|
| Psychiatric | 45 | 43 | 47 | 46 | Insomnia |
| Nervous sy... | 44 | 47 | 42 | 43 | Headache |
| Cardiovas... | 50 | 45 | 49 | 42 | Angina |
| Gastrointe... | 46 | 41 | 52 | 50 | nausea |
| Dermatolo... | 43 | 50 | 51 | 47 | rash |
| Endocrine | 45 | 51 | 41 | 42 | Hyposper... |
| Hematologic | 52 | 43 | 47 | 41 | Anemia |
| Psychiatric | 42 | 47 | 51 | 45 | nervousne... |
| Nervous sy... | 45 | 42 | 47 | 44 | paraesthe... |
| Cardiovas... | 41 | 48 | 47 | 42 | chest pain |
| Gastrointe... | 42 | 45 | 41 | 48 | Abdominal... |
| Dermatolo... | 51 | 54 | 46 | 42 | purpura |
| Hematologic | 53 | 45 | 43 | 32 | leucopenia |
| Psychiatric | 45 | 56 | 48 | 43 | agitation |
| Nervous sy... | 47 | 42 | 39 | 40 | Dizziness |

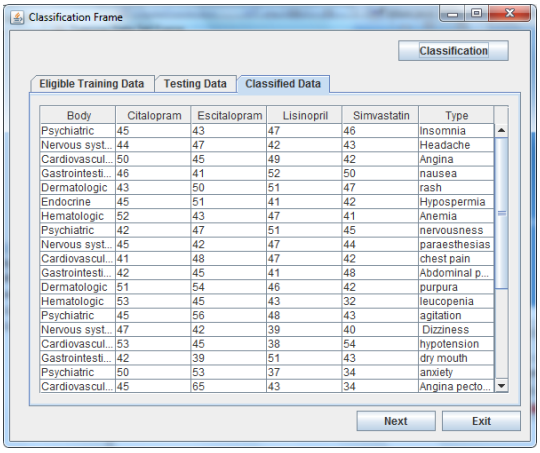
The above diagram specifies training dataset for finding classified data in the testing data. By using training dataset we can find out the type of side effect will occur in body.



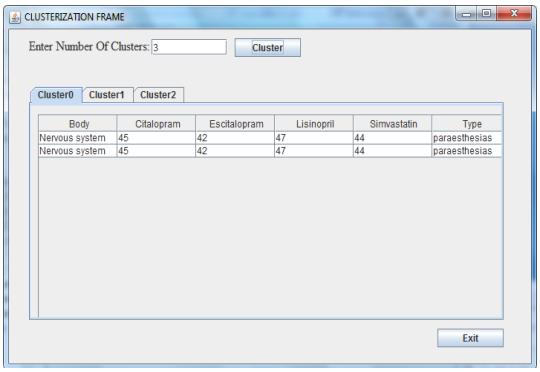

| Body | Citalopram | Escitalopram | Lisinopril | Simvastatin |
|---|---|---|---|---|
| Psychiatric | 45 | 43 | 47 | 46 |
| Nervous system | 44 | 47 | 42 | 43 |
| Cardiovascular | 50 | 45 | 49 | 42 |
| Gastrointestinal | 46 | 41 | 52 | 50 |
| Dermatologic | 43 | 50 | 51 | 47 |
| Endocrine | 45 | 51 | 41 | 42 |
| Hematologic | 52 | 43 | 47 | 41 |
| Psychiatric | 42 | 47 | 51 | 45 |
| Nervous system | 45 | 42 | 47 | 44 |
| Cardiovascular | 41 | 48 | 47 | 42 |
| Gastrointestinal | 42 | 45 | 41 | 48 |
| Dermatologic | 51 | 54 | 46 | 42 |
| Hematologic | 53 | 45 | 43 | 32 |
| Psychiatric | 45 | 56 | 48 | 43 |
| Nervous system | 47 | 42 | 39 | 40 |
| Cardiovascular | 53 | 45 | 38 | 54 |
| Gastrointestinal | 42 | 39 | 51 | 43 |
| Psychiatric | 50 | 53 | 37 | 34 |

The above diagram specifies the type of data to be classified by using training dataset. The type of

data can be specified for the testing data purposeofclassification process.




The above diagram specify for functionality of classification approach. By using training data set we classify the testing data by using naïve Bayes theorem.




The above diagram specifiesclusterization process on side effects of drugs. By using clustering process we group the same type of side effects in to single group and also specify the side effect that will occur in the body.

## V.CONCLUSIONS

We consider the proposed system is mainly focusing on the classification of the type of side effects that will occur by taking over dose of drugs and also to group the same type of side effects into one. By implementing the above discussed concepts we can provide more efficiency on classification and specify which drug will cause which side effect. In this we are performing classification approach by using Naïve Bayesian classifier and clustering of the same type of side effects by using Word Comparator Clustering algorithm.

## VI. REFERENCES

[1] M. Hu and B. Liu, "Mining and summarizing customer reviews,"in Proc. 10th ACM SIGKDD Int. Conf. KDD,Washington, DC, USA,

[2]. B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Ret.*, vol. 2, no. 1–2, pp. 1–135, Jan. 2008.

[3]. L. Zhuang, F. Jing, and X. Zhu, "Movie review mining and summarization,"
in*Proc. 15th ACM CIKM*, New York, NY, USA, 2006,pp. 43–50

[4]. A. Névéol and Z. Lu, "Automatic integration of drug indications from multiple health resources," in *Proc. 1st ACM Int. Health Inform. Symp.*, New York, NY, USA, 2010, pp. 666–673.

[5]. J. Zrebiec and A. Jacobson, "What attracts patients with diabetes to an internet support group? A 21-month longitudinal website stuey," *Diabetic Med.*, vol. 18, no. 2, pp. 154–158, 2008.

[6] J.C. Benaloh. Secret sharing homomorphisms: Keeping shares of a secretsecret. In *Crypto*, pages 251–260, 1986.

[7] J. Brickell and V. Shmatikov.Privacy-preserving graph algorithms inthe semi-honest model. In *ASIACRYPT*, pages 236–252, 2005.

[8] D.W.L. Cheung, J. Han, V.T.Y. Ng, A.W.C. Fu, and Y. Fu. A fastdistributed algorithm for mining association rules. In *PDIS*, pages 31–42, 1996.

[9] D.W.L Cheung, V.T.Y. Ng, A.W.C. Fu, and Y. Fu. Efficient miningof association rules in distibuted
databases. *IEEE Trans. Knowl. DataEng.*, 8(6):911–922,

[10]. S. Lacoste-Julien, F. Sha, andM. Jordan, "DiscLDA: Discriminativelearning for dimensionality reduction and classification," in *Proc.Adv. NIPS*, 2008, pp. 897–904.

[11]. J. Demšar, "Statistical comparisons of classifiers over multipledata sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Jan. 2006.

[12].R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th Int. Conf. VLDB*, San Francisco, CA, USA,1994, pp. 487–499.

[13]. A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Proc. Conf. Human Lang. Technol. Emp. Meth. NLP*, Stroudsburg, PA, USA, 2005, pp. 339–346.

[14]. B. Liu, M. Hu, and J. Cheng, "Opinion observer: Analyzing and comaring opinions on the web," in *Proc. 14th Int. Conf. WWW*, New York, NY, USA, 2005, pp. 342–351.

[15]. S. Baccianella, A. Esuli, and F. Sebastiani, "Multi-facet rating of product reviews," in *Proc. 31st ECIR* , Berlin,, Germany, 2009, pp. 461–472.

[16]. D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.

[17]. Y. Jo and A. Oh, "Aspect and sentiment unification model for online review analysis," in *Proc. 4th ACM Int. Conf. WSDM*, New York, NY, USA, 2011, pp. 815–824.

[18]. D. Blei and J. Mcauliffe, "Supervised topic models," in *Proc. Adv. NIPS*, 2007, pp. 121–128.

[19]. D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *Proc. Adv. NIPS*, 2003.

[20] H. Lee, J. Yoo, and S. Choi, "Semi-supervised nonnegative matrix factorization," *IEEE Signal Process. Lett.*, vol. 17, no. 1, pp. 4–7, Jan. 2010.

## BIOGRAPHIES:

Haritha Paidi is student in M.Tech (CSE) in Dadi Institute of Engineering & Technology, Anakapalle, Visakhapatnam.She has receivedher B.tech(C.S.E) from Visakha Institute of Engineering and Technology, Visakhapatnam. Her

interesting areas are Data Mining and network security.

L. Prasanna Kumar is working as Assoc. Professor in Dadi Institute of Engineering & Technology, Visakhapatnam, Andhra Pradesh.. His research areas include Big Data, Data Mining, Artificial intelligence.