A framework for predicting the effectiveness of drugs based on drug reviews

Dr. S. Adaekalavan
Assistant Professor
Department of Computer Science,
J.J. College of Arts and Science (Autonomous)
Pudukkottai, India
kingsmakers@gmail.com

***Abstract*—**This paper proposes a framework for predicting the effectiveness of the drugs used by the patients based on the reviews given by the patients. This framework consists of five phases to draw the predictions. The reviews given by the patients will be a free form text. This leads to pre process the text for the classification and prediction model generation. The framework uses a series of text preprocessing methods to identify the valuable terms in the reviews. The word net database is used to identify the terms. The clusters are formed based on the terms identified. Based on the clusters, a classification model is generated. This model is then used to predict the effectiveness of the drugs. In addition to prediction, the framework also presents the statistical reports about the clusters formed. The effectiveness of the framework is also analyzed in this paper.

Keywords—Clustering, Classification, Prediction, Drug Reviews, Text Preprocessing

I. INTRODUCTION

One of the critical stages in healthcare domain is to predict the effectiveness of the drugs prescribed by the doctors. This process is twofold. It should be done at the time of clinical trials [2]. Again it should be done after the usage by the medical practitioners. An emerging discipline in systems biology is Systems medicine which is aimed to integrate clinical databases with large-scale molecular interaction data to explicate about drugs prescribed for diseases [1]. This paper focuses on the second aspect. Personalized medicine refers to the customized therapy for an individual patient rather than the approach adopted for set of patients. Drug sensitivity prediction plays a key role in personalized medicine. The idea of personalized medicine is not new approach. It has been practiced since the time of Hippocrates [13]. They used to treat the patients based on their bodily fluids.

Usually the patients are requested to record their feedback about the drugs used by them. Earlier it seems to be a hectic process. But now a day's with the existence of social media it becomes quite common and easy. Sentiment analysis is the computational study of people's attitudes, appraisals, and opinions about individuals, issues, entities, topics, events, and products as well as their attributes [3]-[11]. Though Sentiment Analysis is having a wide range of real time applications, it is technically challenging [12]. This work considers the reviews by the patients as a primary source for the predictive model.

The people are getting more aware because of the emerging technological developments has led to the growth of large amount of data. This enormous data contain more valuable knowledge which will be useful in making important decisions. Data mining is a major procedure of distinguishing valid, novel, potential, valuable and eventually rational designs in information [14]. Cluster Analysis is one among the most commonly used strategy in data mining. It may be either used as a solitary instrument for understanding the knowledge base there by helps in decision making or as a pre-processing scheme for other data mining techniques. The proposed framework uses the hierarchical clustering for the classification model generation.

This research paper is organized as follows. The review of the literature is presented in the section II. The proposed framework is explained in the section III. The experimental results are analyzed in section IV. The conclusion and future directions are recorded in section V.

I. BACKGROUND STUDY

This research work is motivated by several works. This section describes the background study of the proposed framework. Databases are the basic component needed to analyze large volume of data efficiently. Data mining algorithms on are instrumental and boosting the ability to analyze data significantly. In order to perform data analysis, data integrity and management considerations are inevitable. The overview for the data mining techniques and their applicability in the database perspective can be found in [20]. Clustering is the one of the most promising Data mining techniques. In [21], the authors explained about the most efficient K-means clustering.

Nearest neighbor classifier is a nearness-based classifier which use distance-based measures to perform the classification. The main idea behind the grouping is that documents which belong to the same class are more likely "similar" or close to each other based on the similarity measures. One such measure used for finding the similarity is cosine measure. The grouping of the test document is incidental from the cluster labels of similar documents in the training set. If we consider the k-nearest neighbor in the training data set, the approach is called k-nearest neighbor classification and the most common class from these k neighbors is reported as the class label [21].

Information extraction is the task of extracting structured information from unstructured text in a automatic fashion. In a biomedical domain, unstructured text may comprises of prescriptions, discharge summaries, reviews by the patients,

scientific articles in biomedical literature and medical information found in clinical information systems. Information extraction is typically considered as a preliminary step dealing with the preprocessing activity in other text mining applications such as question answering [18], hypothesis generation[19] and summarization.

A. *Dataset Description*

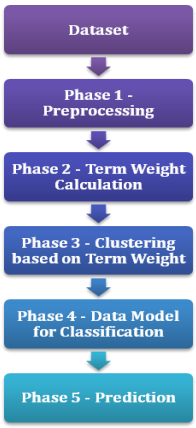
The dataset used for this proposed research work is downloaded from the UCI repository. The name of the dataset is Drug Review Dataset [17]. It consists of 8 attributes and 4143 instances. The dataset provides patient reviews on specific drugs along with related conditions. Furthermore, reviews are grouped into reports on the three aspects benefits review, side effects and overall comment. The data is split into two partitions namely, a train (75%) consists of 3107 instances and a test (25%) consists of 1036 instances. The attributes of the dataset are shown in the following Table I.

TABLE I. ATTRIBUTES INFORMATION OF DRUG REVIEW DATASET

Attribute Information			
#	***Attribute Name***	***Type***	***Attribute Description***
1	urlDrugName	categorical	name of drug
2	condition	categorical	name of condition
3	benefitsReview	text	patient on benefits
4	sideEffectsReview	text	patient on side effects
5	commentsReview	text	overall patient comment
6	rating	numerical	10 star patient rating
7	sideEffects	categorical	5 step side effect rating
8	effectiveness	categorical	5 step effectiveness rating

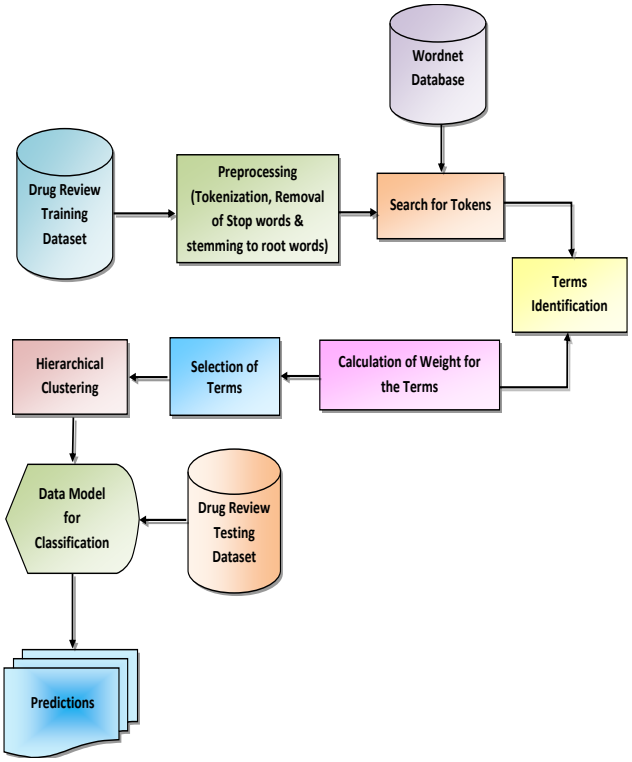
II. PROPOSED FRAMEWORK

The proposed framework consists of five major phases. The work flow of the five phases is depicted in figure 1. The phases are explained in the following sections. In order to get more clear view about the proposed framework, the architecture diagram is also presented.



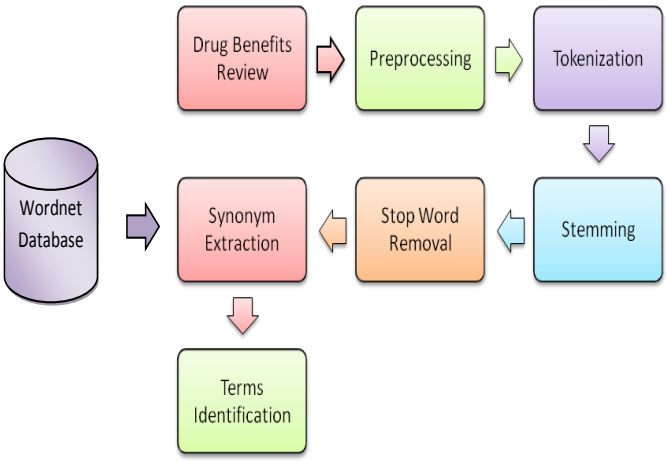
Phases of the proposed Framework

The detailed architecture of the proposed framework for predicting effectiveness of drugs based on patients drug reviews is shown in the following figure 2. This architecture encompasses of both training phase and the testing phase.



A. *Phase I - Reviews Preprocessing*

Preprocessing of the Drug benefits reviews is the first phase. This phase plays a key role in predicting the effectiveness of the drugs prescribed. The preprocessing of the drug reviews includes Tokenization, Stemming and Stop Word Removal. Once the reviews are preprocessed, the terms representing the opinions are extracted using the Wordnet database. In addition to the exact term matches the synonyms also extracted for better clustering.



The steps involved in the drug review preprocessing are explained in this section. Let us consider the review *"I think that the Lyrica was starting to help with the pain, but the side-effects were just too severe to continue"* to elucidate the steps involved in preprocessing.

1) Tokenization: Tokenization is the process of breaking a stream of textual content up into words, terms, symbols, or some other meaningful elements called tokens.

After Tokenization: *I, think, that, the, Lyrica, was, starting, to, help, with, the, pain, but, the, side, effects, were, just, too, severe, to, continue*

2) Stemming: Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. Here the Porter's algorithm [15] is used for Stemming.

After Stemming: *I, think, that, the, Lyrica, was, start, to, help, with, the, pain, but, the, side, effect, were, just, too, severe, to, continue*

3) Stop Word Removal: A stop word is a commonly used word (such as "the", "a", "an", "in") that should be ignored for searching and createing clustering.

After Stop words Removal: *think, start, help, pain, side, effect, severe, continue.*

4) Synonym Extraction: The Wordnet database is searched for the occurence of the words identified after stop word removal process. If there is no exact match found, then synonym of the words will be extracted otherwise this step will be skipped. Here the opinion words alone are considered as terms.

After Synonym Extraction: *help* is associated with the keyword *pain, severe* is associated with the keyword *side effect.*

5) Terms identification: The words will be checked for the type of opinion. The opinion type may be "Positive", "Negative" or "Neutral". The terms are identified in this step.

After Terms identification: *help* is *positive opinion, severe* is *negative opinion.*

B. Phase II - Term Weight Calculation

The weight for the terms identified in Phase I are calculated in this phase. Based number of occurrences of the terms, number of positive terms, number of negative terms and number of neutral terms, the weight of the terms for the drug review is calculated. The weights are calculated based on the research work [16].

$$w_i = 0.30 * nocc_i + 0.40 * npos_i + 0.20 * nneg_i + 0.10 * nneu_i$$

where, w_i is the weight of the review i, $nocc_i$ is number of occurrences of the terms, $npos_i$ is the number of positive terms, $nneg_i$ is the number of negative terms and $nneu_i$ is the number of neutral terms in review i. The weight is distributed differently for each category. The positive opinions will be given greater weight when compared to the other categories. The count values for the parameters and the weight measures are shown in Table II.

Count and Weight for the terms

Parameter	Values	Weight
$nocc_i$	2	0.20
$npos_i$	1	0.40
$nneg_i$	1	0.20
$nneu_i$	0	0.10

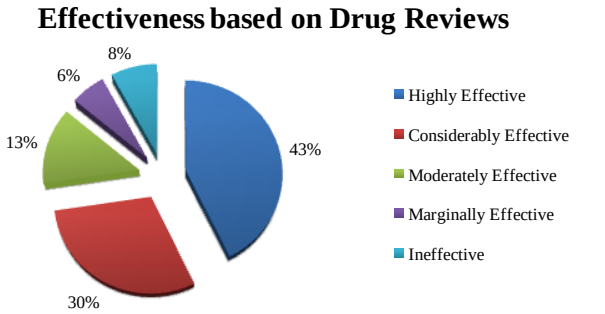
$$w_i = 0.30 * 2 + 0.40 * 1 + 0.20 * 1 + 0.10 * 0 \ = 1.2$$

C. Phase III - Clustering based on Term Weights

Once the weights are calculated for all the drug reviews in the training dataset, the dataset is clustered using the hierarchical clustering algorithm. The weights are used as the measure for the cluster formation. Five clusters are formed based on the weight measure. Depending upon the values, top 20% weights will be categorized as Highly effective, next 20% as considerably effective, next 20% as moderately effective, next 20% as marginally effective and the last 20% as ineffective.

D. Phase IV - Data model for Classification

The data model for classification and prediction is generated in this phase. The model will accept the testing data, find the term weight of the drug review, identify the cluster and predict the effectiveness of the drugs based on the review. The training data is also analyzed by the statistical reports about the drug effectiveness in this phase. They are depicted in figure 4.



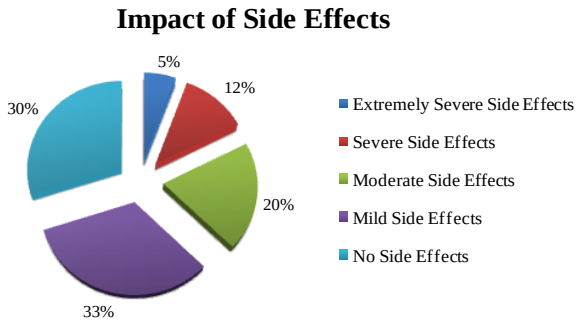


Fig. 4(a)

Fig. 4(b)

Fig. 1. Statistical Reports on the Training dataset

E. Phase V - Predication

This phase is used to draw predictions from the classification model generated in Phase IV. The performance of the proposed framework can be evaluated with the help of this phase. Here the testing dataset consists of 1036 records.

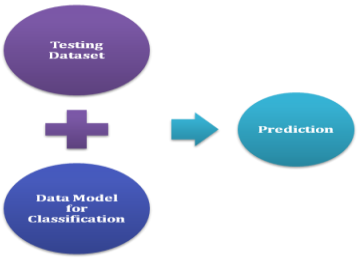


Fig. 2. Phase V - Prediction process

III. RESULTS AND DISCUSSION

The performance of the proposed framework to predict the effectiveness of drugs based on reviews is analyzed in this section. The metrics used for evaluation are Accuracy, False Negative Rate, Sensitivity, Specificity and Precision. The performance of the proposed method is compared with that of other methods like SVM, Naive Bayes and Decision Tree algorithm are discussed below.

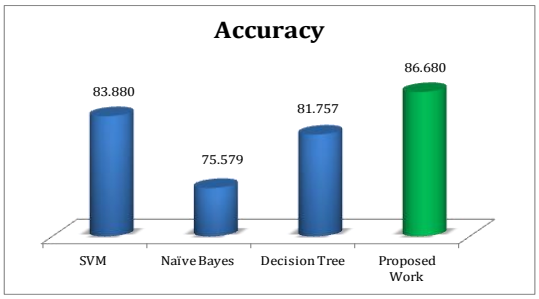


Fig. 6. (a) Performance in terms of Accuracy

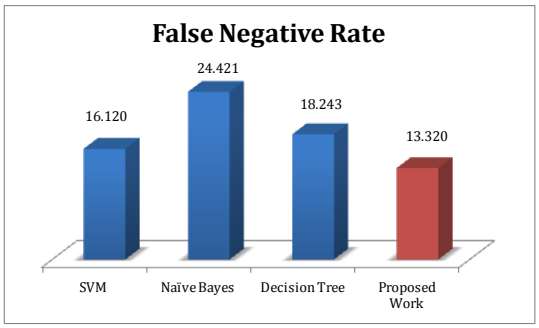


Fig. 6. (b) Performance in terms of False Negative Rate

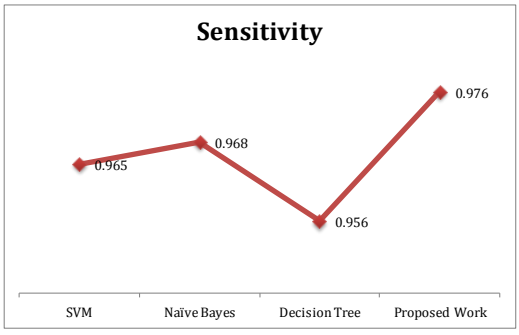


Fig. 6. (c) Performance in terms of Sensitivity

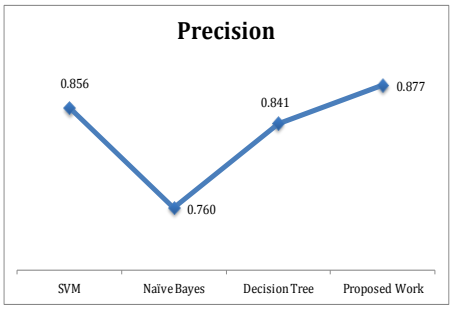


Fig. 6. (d) Performance in terms of Specificity

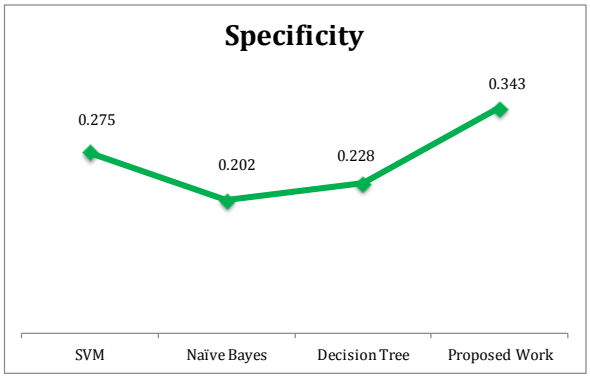


Fig. 6. (e) Performance in terms of Precision

Fig. 3. (a)-(e) Performance Analysis of the proposed Framework

From the above figures 6.(a)-6.(e), it is evident that the proposed framework yields better results when compared to the existing methods. The computational complexity of the proposed work is comparatively high than the existing methods. This is primarily due to the preprocessing of the drug benefits review.

IV. CONCLUSION

This paper proposes a new framework to predict the effectiveness of drugs based on the patients benefits review. The proposed framework first identifies the terms in the patients review about the drugs being used. This requires preprocessing of the dataset. The term based weight and polarity of the opinions are retrieved during the next phases. The training dataset is clustered and the classification model is generated for predictions. The statistical reports of the proposed framework is explained in the previous section. The proposed work is evaluated with the help of testing dataset. Based on the performance analysis the proposed work gives 86.68% of accuracy and 13.32 as the False negative ration. The proposed framework offers better performance in terms of specificity, sensitivity and precision.

REFERENCES

Lamb, J., Crawford, E.D., Peck, D., et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313, 2006, pp. 1929–1935.

Nir Atias and Roded Sharan, An Algorithmic Framework for Predicting Side Effects of Drugs, Journal of Computational Biology, Volume 18, Number 3, 2011, pp. 207–218, DOI: 10.1089/ cmb.2010.0255

A. Immonen, P. Pääkkönen, and E. Ovaska, Evaluating the Quality of Social Media Data in Big Data Architecture, IEEE Access, vol. 3, pp. 2028-2043, 2015.

D. Jiang, X. Luo, J. Xuan, and Z. Xu, Sentiment computing for the news event based on the social media big data IEEE Access, vol. 5, pp. 2373-2382, 2016.

M. N. Injadat, F. Salo, and A. B. Nassif, Data mining techniques in social media: A survey, Neurocomputing, vol. 214, pp. 654-670, Nov. 2016.

T. A. A. Al-Moslmi, Machine Learning and Lexicon-Based Approach for Arabic Sentiment Analysis. Bangi, Malaysia: Fakulti Teknologi & Sains Maklumat/Institut, 2014.

N. Omar, M. Albared, T. Al-Moslmi, and A. Al-Shabi, A comparative study of feature selection and machine learning algorithms for arabic sentiment classification, in Information Retrieval Technology. Springer, 2014, pp. 429-443.

M. Bouazizi and T. Ohtsuki, A pattern-based approach for sarcasm detection on twitter, IEEE Access, vol. 4, pp. 5477-5488, 2016.

F. Bertola and V. Patti, Ontology-based affective models to organize artworks in the social semanticWeb, Inf. Process. Manage., vol. 52, no. 1, pp. 139-162, 2016.

G. Vinodhini and R. M. Chandrasekaran, A sampling based sentiment mining approach for e-commerce applications, Inf. Process. Manag., vol. 53, no. 1, pp. 223-236, 2016.

R. Piryani, D. Madhavi, and V. K. Singh, Analytical mapping of opinion mining and sentiment analysis research during 2000 - 2015, Inf. Process. Manag., vol. 53, no. 1, pp. 122-150, 2016.

Tareq Al-Moslmi, Nazlia Omar, Salwani Abdullah, and Mohammed Albared, Approaches to Cross-Domain Sentiment Analysis: A Systematic Literature Review, IEEE Access, Volume 5, 2017, pp. 16173 - 16192

Steele, F.R. Personalized medicine: Something old, something new. Future Med. 2009, 6, 1–5.

Shivangi Bhardwaj, Data Mining Clustering Techniques – A Review, International Journal of Computer Science and Mobile Computing, Vol. 6, Issue. 5, May 2017, pg.183 – 186

Porter, Martin F. 1980. An algorithm for suffix stripping. *Program* 14 (3): 130-137

Kwang Mong Sim, "Toward an Ontology-Enhanced Information Filtering Agent" in ACM SIGMOD Rec., Vol. 33, Mar 2004

Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In Proceedings of the 2018 International Conference on Digital Health (DH '18). ACM, New York, NY, USA, 121-125.

Sofia J Athenikos and Hyoil Han. 2010. Biomedical question answering: A survey, Computer methods and programs in biomedicine 99, 1 (2010), 1–24.

G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

K. Elissa, "Title of paper if known," unpublished.

R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron