Classification of Drugs Reviews using W-LRSVM Model

Asha S Manek[1], Kailash Pandey K[2], P Deepa Shenoy[2], M. Chandra Mohan[3], Venugopal K R[2]
[1]Research Scholar, Department of Computer Science and Engineering,
[1,3]Jawaharlal Nehru Technological University, Hyderabad, India
[2]University Visvesvaraya College of Engineering, Bangalore University, Bangalore, India

***Abstract*— Opinion mining provided less opportunity to discuss their experiences about drugs so reviewing about it was difficult. Recent findings show that online reviews and blogs on drugs are important for patients, marketers and industries. Collecting the information for drugs from the website and analyzing is a challenge. A model is designed by proposing an algorithm which crawls information from the web to analyze reviews of drugs. Reviews were crawled for five different drugs using the algorithm. The W-Bayesian Logistic Regression and Support Vector Machine (W-LRSVM) model was trained for different split ratios to obtain the accuracy of 97.46%. Experimental results on reviews of five different drugs showed that the proposed model gave better results compared to other classifiers.**

Keywords—Drug Reviews, Crawling, Sentiment Analysis, Split Ratio, Samplings.

I. INTRODUCTION

The rise of social media such as blogs and social networks has fueled to contribute their contents to the internet. Users can share their experience about a particular product via these blogs and social networks. These experiences nothing but reviews are categorized as positive and negative, which help marketers or industries to improve the product and also help other users in reviewing the product.

Sentiment analysis, also called opinion mining is a field of study that analyzes people's opinions or sentiments about the entities such as products, services and organizations[1]. Previous studies of opinion mining provide limitless opportunities for patients to discuss their experiences with drugs. Therefore, even companies get limitless opportunities to receive feedback on their products and services[2-4] because of minority groups of patients on the Internet. Furthermore, recent studies have shown that patient opinions are also useful and important with medical professional opinions[5-8], especially for drugs with afflicting side effects.

Many patients hope to get more information from other patients with similar conditions. They can also share their experience and propose practical ways to alleviate symptoms and side effects of the drugs. These online communities were found to have positive impacts on patient health[9-11]. Sentiments (opinions) are of two types: regular and comparative opinions. Regular opinion is often referred to as a simple opinion and further divided into direct opinion and indirect opinion. A direct opinion is an opinion which is expressed directly on an entity or an entity aspect, whereas indirect opinion is expressed indirectly on an entity or aspect of an entity based on its effects on some other entities. Comparative opinions express a relation of similarities or differences between two or more entities and/or a preference of the opinion holder based on some shared aspects of the entities[1].

These sentiments further classified using supervised learning and unsupervised learning. Supervised learning discovers patterns in data and relates data attributes to class attributes. The values of class attribute for further data instances are then predicted by utilizing these patterns. However, in some applications, there are no class attributes. Then the sentiments can be classified using unsupervised learning also called clustering, which organizes data instances into groups called clusters such that the data instances in same cluster are similar to each other and data instances in different clusters are very different from each other[12].

In this paper, a model is designed and an algorithm is proposed in which the drug reviews are crawled from the web and classified drug reviews as positive and negative. A user interface is designed which helps patients and medical professionals in analyzing the results for a particular drug.

The rest of the paper is organized as follows. Section II covers the related work. Implementation of the proposed model is presented in Section III. In Section IV, we compare the proposed model with other algorithms and accuracy, precision and recall is calculated for each drug. Finally, we conclude in Section V.

II. RELATED WORK

Probabilistic Aspect Based Mining Model (PAMM) deals with aspects related to drugs. PAMM is a supervised algorithm which finds the aspects correlated to one class labels only. It uses all the reviews and finds aspects that are specific to the target class. For a given corpus, it calculates weights for the aspects and classify reviews. The performance is evaluated for top *K* aspects using the Mean Pointwise Mutual Information (Mean PMI) method and classified using SVM

978-1-4673-6540-6/15/$31.00 ©2015 IEEE

with Linear kernel[13].

In understanding a corpus Topic Modelling (e.g., Linear Discriminant Analysis) approach i.e a set of topics, which are represented by multinomial distributions over vocabulary words is inferred. When the words of a topic are sorted according to the probabilities, high probability words on a topic are semantically correlated and the concept or aspect of the topic is manually captured. For example, Joint Sentiment/Topic (JST) model, Aspect and Sentiment Unification Model (ASUM) and Topic Sentiment Mixture (TSM) were proposed to extract both aspects and predict their associated sentiments. Nevertheless, these aspects based opinion mining methods may not be appropriate to address the problem defined in the previous section as extracted aspects may not be related to the specified class labels and the performance depends on selection of seed words which is done manually[14].

Recently, topic modeling with supervised label information has become an interest of research. The supervised LDA (sLDA) proposed by Blei and McAuliffe[15], can handle different forms of supervised information during topic inference.

Ramage et al[16] proposed DiscLDA to process discriminative information and find topics specific to individual classes as well as topics shared across different classes. Labeled LDA is another generalization of LDA. It allows multi-label supervision and associates each label with one topic in direct correspondence.

Apart from probabilistic algorithms, deterministic methods such as Non-Negative Matrix Factorization (NMF) for topic modelling were also proposed in the paper[17]. The data matrix was decomposed into two low rank matrices and topics can be identified.

Semi-Supervised NMF (SSNMF) is an extension proposed recently to incorporate the supervised information into NMF. The topics identified are closely related to the supervised information[18].

An enormous amount of review data is generated everyday in various applications on the web. The traditional association rule mining algorithms were developed to find positive associations between items existing in online web transactions. Analyzing customer behavior patterns as well as finding positive associations between drug reviews can also be possible by using incremental and dynamic association rule mining with genetic algorithm[19-21].

In paper[22], an effective and novel model SentReP is proposed to classify the sentiments of movie reviews using Repetitive Pre-processing technique to obtain tokenized wordlist. SentReP model is tested with K-NN, Naïve Bayes, SVM Linear and SVM Stratified classifiers across different movie review data set with different sizes obtaining the results and performance with SVM Linear algorithm with an accuracy of 97.25%, Precision of 100% and Recall rate of 97.25% for 1700 positive and 1700 negative movie reviews.

III. IMPLEMENTATION

A. Data Source

The first step is to find the most popular website dedicated only to drugs. There are many popular websites like Drugs.com, DrugsLib.com, WebMd etc. The reviews are crawled from Drugs.com. The top drugs are selected from it and reviews collected for 5 drugs: Citalopram, Escitalopram, Lisinopril, Lyrica and Oxcycodone. Collecting the less number of reviews from the website is easier, but for more number (1000's) is difficult and time consuming. So an algorithm is proposed for this task. The algorithm uses Jsoup to parse the web pages and select only the review text avoiding other noisy data. The flowchart for crawling reviews is as shown in Fig 1.

First the drug to be searched is given as an input from a User Interface or an alternative approach is to store the name of the drugs in a database, query it and pass to method. The method selects only the URL from the search page and hit the page for that URL, selects only the review text and move to the next page. The same process is repeated for other URLs. The review data set will be written in a document or text file.

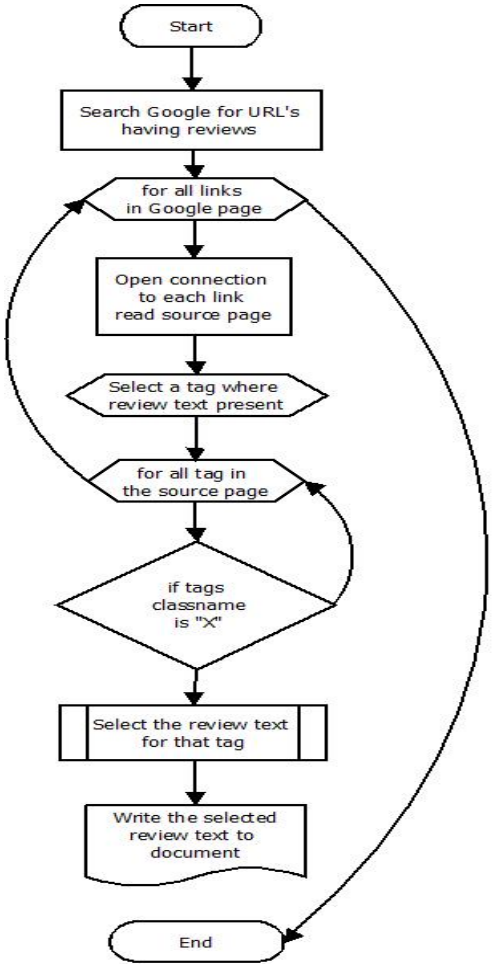


FIG 1. FLOWCHART FOR REVIEWS COLLECTION.

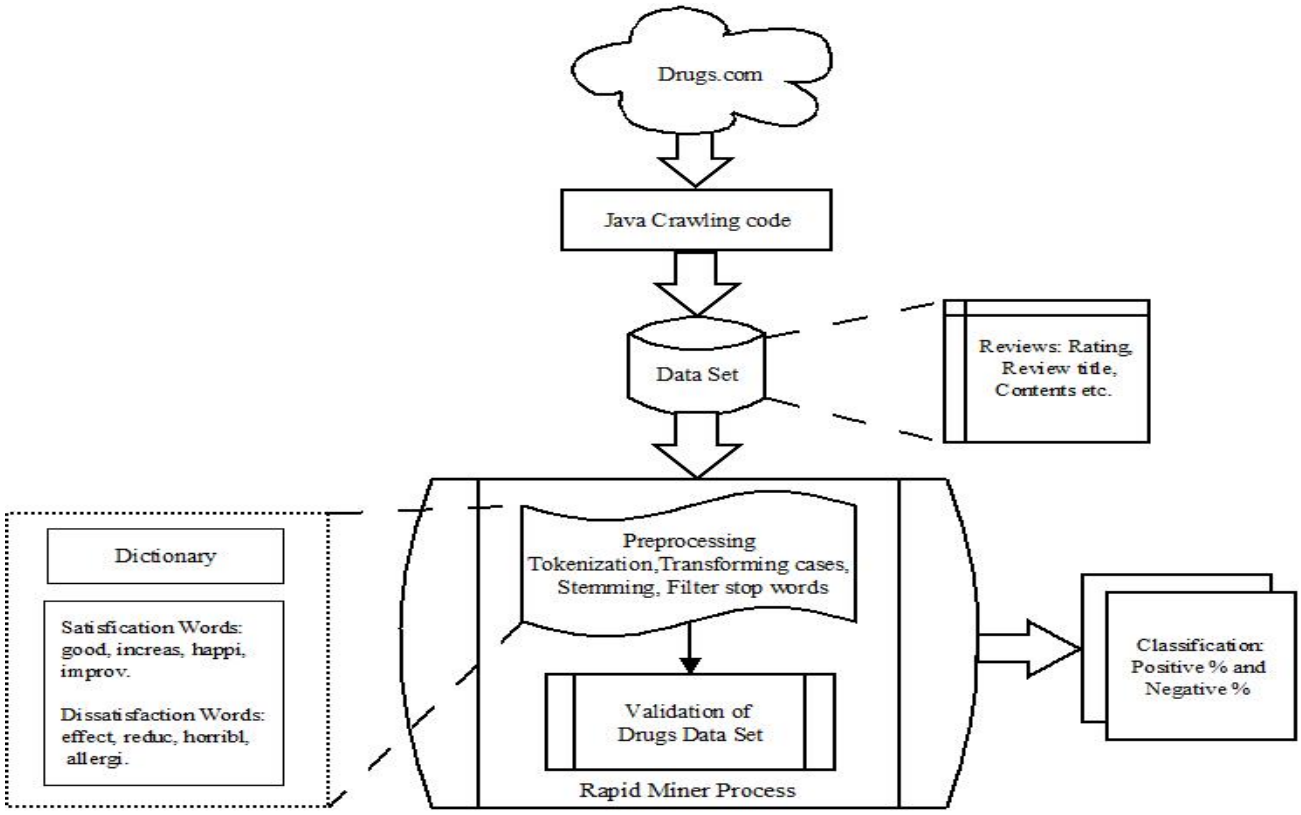


FIG 2. ARCHITECTURE OF W-LRSVM MODEL.

B. *W-Bayesian Logistic Regression Support Vector Machine (W-LRSVM) Model*

W-LRSVM model is designed using RapidMiner tool. Data set collected need to be preprocessed to remove unwanted words and stop words for classification as well as to increase the efficiency of the model. The word list got from the previous step is classified and performance is evaluated for it. The architecture of W-LRSVM is as shown in Fig 2. The steps involved in the model are as follows:

1) Preprocessing:

The data set is a mixture of both positive and negative reviews. In this process, a series of sub tasks are involved as follows: Each reviews is completely read and words occurring in the review undergo tokenization, case transformation, stemming using Porter's algorithm and Snowball stemmer operators and finally all the English stop words present in the reviews are identified and filtered out.

2) Validation:

The validation of the preprocessed data is as follows and shown in Fig 3:

a) Calculation of weights: The weight of attributes is calculated with respect to the label attributes by using correlation. The higher the weight of an attribute, the more relevant it is considered. A correlation is a number between -1 and +1 that measures the degree of association between two attributes.

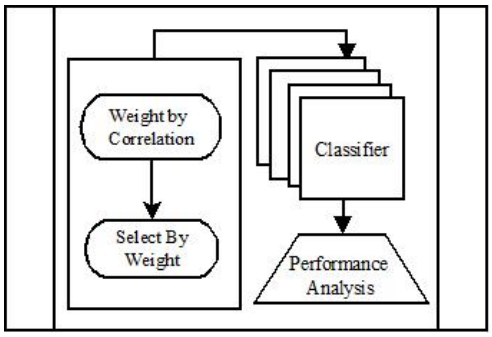


FIG 3. PROCESSES IN VALIDATION OF DRUG DATA SET.

b) Select by weights: Only those attributes is selected whose weight satisfy the specified criterion with respect to the input weights. Only the top *K* (5, 15 and 25) attributes is selected for further evaluation.

c) Performance Analysis: The top *K* words are evaluated using SVM Linear with W-Bayesian Logistic Regression classifier. Accuracy, precision and recall are

calculated for input by varying the split ratio from 0.7 to 0.9 using stratified sampling.

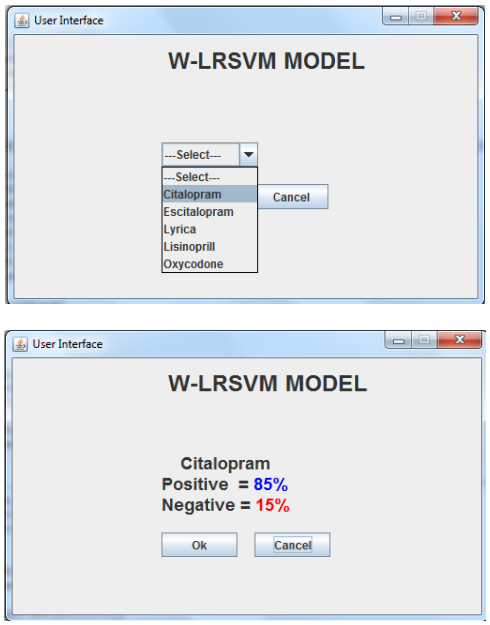


FIG 4. USER INTERFACE OF W-LRSVM MODEL.

In the classsification method, the terms true positive (TP), true negative (TN), false positive (FP), false negative(FN) compare results of the classifier. Positive and negative refers to the clasifiers prediction whereas true and false refers to whether the prediciton made corresponds to the observation. The terms Accuracy, Precision and Recall can be defined as follows:

Accuracy: The effectiveness of classifier is measured by Accuracy. The accuracy is the proportion of both true positives and true negatives among the total number of cases tested.

Accuracy = {True Positive + True Negative} / {True Positive +True Negative +False Positive +False Negative}.

Precision: The exactness of classifier is called as Precision. Less false positive gives higher Precision value.

Precision = True Positive / {True Positive + False Positive.}

Recall: The completeness of a classifier is measured by Recall. Less false negatives gives higher Recall value.

Recall =True Positive / {True Positive + False Negative}.

C. User Interface

The User Interface is designed which consists of a combo box containing drugs list and a description of the drug. The drug selected passed as an argument to the reviews collection algorithm. The output i.e. reviews is passed to the model to display the percentage of positive and negative reviews which can be easily understood by the user. The User Interface is as shown in the Fig 4

IV. RESULTS ANALYSIS

A. W-Bayesian Logistic Regression Support Vector Machine (W-LRSVM) Model Results.

First the reviews are collected from Drugs.com using the algorithm for 5 drugs- Citalopram, Escitalopram, Lisinopril, Lyrica and Oxycodone. Citalopram and Escitalopram is used for depression, Lisinopril is used to regulate blood pressure, Lyrica drug for neuropathic pain and Oxycodone drug is used to treat severe pain.

TABLE I. SUMMARY OF REVIEWS.

Drug Name	Brand	Usage	Reviews Crawled
Citalopram	Celexa	Depression	4211
Escitalopram	Lexapro	Depression	1580
Lyrica	Lyrica	Blood Pressure	2703
Lisinopril	Prinivil	Neuropathic Pain	2524
Oxycodone	Oxecta	Severe Pain	2669

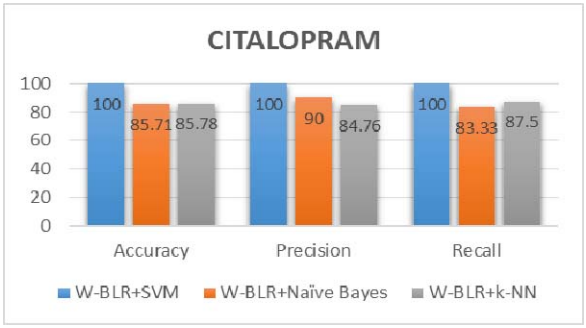
A total of 13,687 reviews is crawled from the website. The summary of reviews is given in Table I.

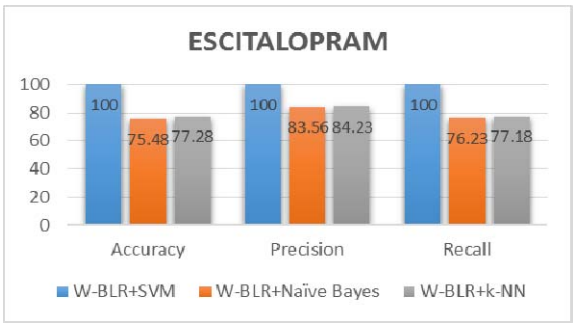
TABLE II. ANALYSIS OF REVIEWS USING THE W - LRSVM MODEL FOR DIFFERENT SPLIT RATIO.

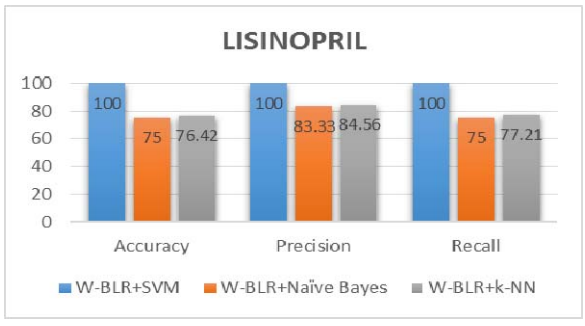
Split ratio	Accuracy %	Error %	Precision %	Recall %
Citalopram (K=27)				
0.7	89.47	10.53	92.31	87.50
0.75	87.50	10.50	90.91	85.71
0.8	84.62	15.38	90.00	80.00
0.85	90.00	10.00	92.86	87.50
0.9	**100.00**	**0.00**	**100.00**	**100.00**
Escitalopram (K=21)				
0.7	50.00	50.00	30.00	37.50
0.75	50.00	50.00	30.00	37.50
0.8	60.00	40.00	50.00	30.00
0.85	33.33	66.67	25.00	25.00
0.9	**100.00**	**0.00**	**100.00**	**100.00**
Lisinopril (K=23)				
0.7	63.64	36.36	80.00	60.00
0.75	77.78	22.22	85.71	75.00
0.8	62.50	37.50	78.71	62.50
0.85	83.33	16.67	87.50	83.33
0.9	**100.00**	**0.00**	**100.00**	**100.00**
Lyrica (K=28)				
0.7	66.67	33.33	33.33	50.00
0.75	55.55	45.55	30.00	42.86
0.8	62.50	37.50	31.25	50.00
0.85	66.67	33.33	33.33	50.00
0.9	**100.00**	**0.00**	**100.00**	**100.00**
Oxycodone (K=20)				
0.7	94.12	5.88	96.88	75.00
0.75	90.00	10.00	92.86	87.50
0.8	94.38	5.62	96.91	75.18
0.85	96.23	3.77	98.32	78.90
0.9	**100.00**	**0.00**	**100.00**	**100.00**

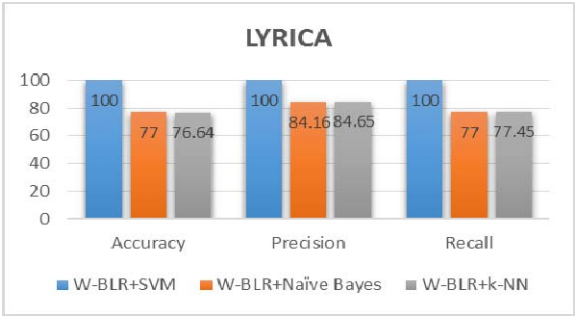
The next step is to classify the reviews. The data set is divided into training and testing data set before classification. The model is trained using the training set and the performance is evaluated for different split ratio starting from 0.7 to 0.9 using stratified sampling method as illustrated in Table II.

The reviews of all drugs were mixed, given as input to W-LRSVM model and performance is compared with different combination of classifiers like W-Bayesian Logistic Regression + Naïve Bayes, W-Bayesian Logistic Regression + k-NN and W-Bayesian Logistic Regression + SVM classifier. The graphical analysis of performance (i.e Accuracy, Precision and Recall) is as shown in Fig 5. With the proposed model, we obtained 97.46% Accuracy, 98.32% Precision and 87.50% Recall for all 5 drugs chosen. Table III shows comparison of mixed reviews of 13,687 reviews with different classifiers for 0.9 split ratio.









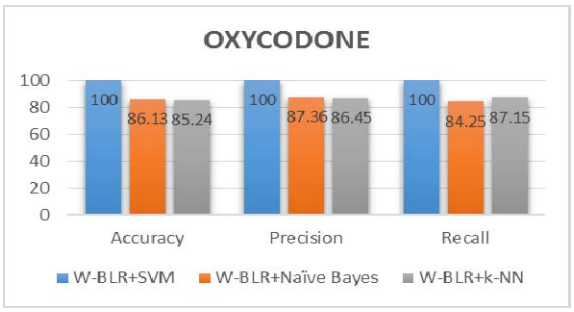


FIG 5. COMPARISON OF ALL DRUGS WITH OTHER CLASSIFIER ENSEMBLE FOR 0.9 SPLIT RATIO.

The W-BLR + SVM gave 97.46% Accuracy, W-BLR + Naïve Bayes showed 76.19% Accuracy and 77.24% Accuracy with W-BLR + k-NN classifier.

The categorization of words for all 5 drugs is as shown in Table V. The categorization of words is done based on usefulness, information, customer experiences and side effects into satisfaction and dissatisfaction words of that particular drug. Some of the satisfaction words are antidepress, work, relief, experience, control, free, etc. Some of dissatisfaction words are bad, stop, horribl, damage, problem, poor, suffer, etc.

TABLE III. COMPARISON OF MIXED REVIEWS WITH OTHER CLASSIFIER ENSEMBLE FOR 0.9 SPLIT RATIO.

Classifier Ensemble	Accuracy %	Error %	Precision %	Recall %
W-BLR + Naive Bayes	76.19	23.81	86.84	64.29
W-BLR + k-NN	77.24	22.76	73.56	75.00
W-BLR + SVM	**97.46**	**2.54**	**98.32**	**87.50**

B. Performance Evaluation

1) Comparison of W-LRSVM Model with reference[23]

Table IV gives the performance evaluation of the proposed model against the work[23]. In[23], Support Vector Machine (SVM) is used to classify the reviews obtaining 80.8% Accuracy whereas W-LRSVM model outperformed obtaining 97.46% Accuracy.

TABLE IV. COMPARISON OF W-LRSVM WITH REFERENCE[23] AND [24].

Classifier	Accuracy %	Precision %	Recall %
W-LRSM	**97.46**	**98.32**	87.50
Compared with Reference[23]	80.8	93.9	**91.7**
Compared with Reference[24]	87.8	-	-

2) Comparison of W-LRSVM Model with reference[24]

The performance of the proposed model is compared with work[24]. Transductive Support Vector Machine (TSVM) is used to classify the data in reference[24], which reported an Accuracy of 87.8%. With a classifier ensemble W-Bayesian Logistic Regression + Support Vector Machine, an Accuracy of 97.46% is achieved as shown in Table IV.

TABLE V. CATEGORIZATION OF WORDS FOR ALL DRUGS.

Satisfaction words of Citalopram		**Dissatisfaction Words of Citalopram**	
recommend	reduc	bad	stress
antidepress	great	sick	attack
life	took	stop	problem
work	felt	reaction	chang
sleep	improv	effect	wors
Satisfaction words of Escitalopram		**Dissatisfaction Words of Escitalopram**	
lot	feel	anxieti	wors
made	work	depress	effect
good	took	side	horribl
immedi	felt	made	suffer
happi	improv	attack	Bad
Satisfaction words of Lisinopril		**Dissatisfaction Words of Lisinopril**	
pressur	experienc	effect	discontinu
good	work	headach	switch
control	reduc	pain	discomfort
lower	great	problem	allergi
pleas	wonder	horribl	difficulti
Satisfaction words of Lyrica		**Dissatisfaction Words of Lyrica**	
feel	work	suffer	weight
relief	sleep	increas	sever
thank	good	terribl	stop
wonder	happi	anymor	bad
great	reduc	damag	horribl
Satisfaction words of Oxycodone		**Dissatisfaction Words of Oxycodone**	
relief	great	lower	increas
lower	benefit	suffer	poor
work	pain	effect	refus
free	control	damag	faint
good	worth	problem	hate

V. CONCLUSION

In this paper, we have proposed novel model W-LRSVM to classify sentiments of drugs reviews. The proposed model is tested with W-BLR+K-NN, W-BLR+Naive Bayes and W-BLR+SVM classifiers. The results and performance analysis shows best performance with W-BLR+SVM classifier ensemble with an accuracy of 97.46%, Precision of 98.32% and 87.50% Recall for mixed review set of drugs crawled from the web. The proposed model works efficiently for large data set of reviews crawled using a review collection algorithm. A user interface is also designed for the W - LRSVM model. Thus the model can be used for any drug reviews analysis.

REFERENCES

[1] Bing Liu,"Sentiment Analysis and Opinion Mining", *liub@cs.uic.edu, April 22,2012*.

[2] A. Ochoa, A. Hernandez, L. Cruz, J. Ponce, F. Montes, L. Li, andL. Janacek, "Artificial societies and social simulation using ant colony,particle swarm optimization and cultural algorithms," in New Achieve-ments in Evolutionary Computation, P. Korosec, Ed. Rijeka, Croatia:Intech, pp. 267–297, 2010.

[3] Cornell and W. Cornell. (2013). How Data Mining Drives Pharma: Information as a Raw Material and Product [Webinar]. [Online]. Available:http://acswebinars.org/big-data W.

[4] L. Toldo, "Text mining fundamentals for business analytics," presented at the 11th Annu. Text Soc. Analytics Summit, Boston, MA, USA, 2013.

[5] J. Sarasohn-Kahn, "The wisdom of patients: Health care meetsonline social media," California Healthcare Foundation, Tech.Rep., 2009.

[6] K. Denecke and W. Nejdl, "How valuable is medical social media data? content analysis of the medical web," *J. Inform. Sci*., vol. 179, no. 12, pp. 1870–1880, 2009.

[7] X. Ma, G. Chen, and J. Xiao, "Analysis on an online health socialnetwork," in Proc. 1st ACM Int. Health Inform. Symp., New York, NY, USA, pp. 297–306, 2010.

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

[9] J. Leimeister, K. Schweizer, S. Leimeister, and H. Krcmar, "Do virtual communities matter for the social support of patients? Antecedents and effects of virtual relationships in online communities," *Inform. Technol. People*, vol. 21, no. 4, pp. 350–374, 2008.

[10] R. Schraefel, R. White, P. André, and D. Tan, "Investigating web search strategies and forum use to support diet and weight loss," in *Proc. 27th CHI EA*, New York, NY, USA, pp. 3829–3834, 2009.

[11] 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.

[12] Bing Liu,"Web Data Mining", liub@cs.uic.edu, Springer-Verlag Berlin Heidelberg 2007.

[13] Victor C. Cheng, C.H.C. Leung, Jiming Liu, and Alfredo Milani, "Probabilistic Aspect Mining Model for Drug Reviews", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 8, August 2014 .

[14] D. Blei and J. Lafferty, "Correlated topic models," in *Proc. Adv. NIPS*, 2006, pp. 147–154.

[15] D. Mimno and A. McCallum, "Topic models conditioned on arbitary features with Dirichlet-multinomial regression," in *Proc.* 24th Conf. Uncertain. Artif. Intell, pp. 411–418, 2008.

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

[17] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proc. 26th Annu. Int. ACM* SIGIR Conf. Res. Develop. Inform. Ret., New York, NY, USA, pp. 267–273, 2003.

[18] 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.

[19] P. Deepa Shenoy, K.G. Srinivasa, K.R. Venugopal, Lalit M. Patnaik. "Dynamic Association Rule Mining using Genetic Algorithms" Journal Intelligent Data Analysis, Volume 9, Number 5, pp. 439-453, 2005.

[20] Shenoy, P. Deepa, K. G. Srinivasa, K. R. Venugopal, and Lalit M. Patnaik. "Evolutionary approach for mining association rules on dynamic databases." In Advances in knowledge discovery and data mining,. Springer Berlin Heidelberg, pp. 325-336, 2003.

[21] Srinivasa, K. G., Karthik Sridharan, P. Deepa Shenoy, K. R. Venugopal, and Lalit M. Patnaik. "A dynamic migration model for self-adaptive genetic algorithms." In Intelligent Data Engineering and Automated Learning-IDEAL 2005, Springer Berlin Heidelberg, pp. 555-562, 2005.

[22] Asha S Manek, Pallavi R P, Veena H Bhat, P Deepa Shenoy, Venugopal K R, M Chandramohan, L M Patnaik, "SentReP: Sentiment Classification of Movie Reviews using Efficient Repetitive Pre-Processing", International Conference IEEE TENCON 2013, 22-25 October 2013, Xi'an China,, ISBN No. 978-1-4799-2825-5, pp. 1-5.

[23] Liu, Xiao, and Hsinchun Chen. "AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums" Smart Health. Springer Berlin Heidelberg, 2013. pp 134-150.

[24] Abeed Sarker , Graciela Gonzalez, "Portable automatic text classification for adverse drug reaction detection via multi-corpus training," Journal of Biomedical Informatics 53 (2015) pp 196–207.