International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

An Effective Stratified K-Fold Algorithm with Logistic Regression for Drug Feedback Data

D. Naga Swathi, Kumaran.U

Abstract: ***Drug reviews are commonly used in pharmaceutical industry to improve the medications given to patients. Generally, drug review contains details of drug name, usage, ratings and comments by the patients. However, these reviews are not clean, and there is a need to improve the cleanness of the review so that they can be benefited for both pharmacists and patients. To do this, we propose a new approach that includes different steps. First, we add extra parameters in the review data by applying VADER sentimental analysis to clean the review data. Then, we apply different machine learning algorithms, namely linear SVC, logistic regression, SVM, random forest, and Naive Bayes on the drug review specify dataset names. However, we found that the accuracy of these algorithms for these datasets is limited. To improve this, we apply stratified K-fold algorithm in combination with Logistic regression. With this approach, the accuracy is increased to 96%.***

Keywords: Vader (Valence Aware Dictionary for sentiment Reasoning), stratified K-fold, Machine Learning (ML), Drug Reviews, Natural Language Processing(NLP)

I. INTRODUCTION

Nowadays, reviews are common for any product we purchase through online or offline mode. Reviews may include comments, tweets, posts, and ratings. These reviews have significance, especially in the fields like medicine, electronic products, etc. However, drug reviews play a predominant role in the field of medicine. These drug reviews help the Food and Drug Administration (FDA) for drug evaluation [18].

Moreover, these drug reviews have popular demand in most of the developed countries. Especially, in countries like United States, people purchase drugs and share their reviews along with the ratings. Pharmacists use this information for better medications [20]. These reviews include the side effects after using the drug, purpose of their purchasing and what was their opinion after using the drug. In this work, we attempt to improve the effectiveness of the reviews given by the patients. The reviews are typically noisy, unintelligible, huge in volume and may generated from a hodge-podge of various sources [14].

Everyone knows that the online reviews are either in unstructured or semi-structured data format, and it is necessary to convert them into structured data. To achieve this, we use different NLP techniques like Word-sense-disambiguate, Sentiment analysis, Text classification, Name-entity recognition, Part-of-speech tagging, Machine translation, Stemming and so on [17]. Natural language processing is a branch of artificial intelligence (AI) that can understand the human language as well as teach machines to understand how to communicate [16]. One of the main applications is language translation such as ad google, check for grammatical accuracy of the text, etc. In this paper we have used sentimental analysis technique to extract and modify the unstructured data into structured data [19].

Sentiment analysis is a part of Natural language processing (NLP) which is used to find the sentiments of the user reviews. Opinions and reviews play a significant role in user satisfaction about a particular entity [8]. Aspect based opinion mining is an integral part of the sentiment analysis in the natural language processing technique. Basically, Aspect based opinion mining is used to analyze whether the text is positive or negative [15]. These approaches recognize sentiment terms and patterns of sentiment expressions in natural language processing by matching the sentiment [13].

As a result of applying NLP Techniques, desired feature extraction can be done successfully and then Machine learning techniques can be applied to the reviews. Machine learning is an application that produces a function to predict future events and able to provide the targets for any new inputs compare with the output and find the error to modify the model accordingly. The primary aim is to learn the computer without human intervention automatically [9]. Various Machine learning algorithms are K-nearest Neighbor, Support Vector Machine, Random Forest, Naïve Bayes, Logistic regression, Decision Tree, Linear regression [12]. In this paper, we applied the VADER (Valence Aware Dictionary for sentiment Reasoning), sentimental analysis on the dataset available in UCI repository and Kaggle datasets to perform the drug reviews.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.

* Correspondence Author
D.Naga Swathi*, Department of Student Computer Science and Engineering, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India. Email: nagaswathidammalapati@gmail.com
Kumaran.U, Department of Computer Science and Engineering, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India. Email : u_kumaran@blr.amrita.edu

We feed this data as inputs to various machine learning algorithms to find the accuracy of the data, there are useful for the pharmaceutical industries to improve the dosage, packing's and chemical compositions based on the reviews, which can minimize the side effects and improve the effectiveness of the drug.

To improve further, we have applied the stratified algorithm combined with the logistic regression technique.

II. LITERATURE SURVEY

Liu J et al. [19] have proposed an approach which uses grassroots data from the internet, accessing the drug review system and collecting data from users and reviewing latent information by searching and inquiring. The web-based multi-model spoken dialog program has been introduced to allow users from different websites to know the medication efficiency and side effects. This system provides an evaluation of speech recognition, parse coverage and reaction to the test. Allowing the user to share their experience for user-generated content through speech and text and the other source such as universal speech interface system [1]. Cheng VC et al. [2] have proposed a new technique called Probabilistic Aspect Mining Model (PAMM) for mining is a unique feature here, where PAMM classifies just one class instead of deriving all the class features. Estimation of this is not complex as one matrix must be taken from the data of training. Cheng V et al. proposed a new approach which is extension of the probabilistic and unregulated method. It can identify patient sentiment reviews before using PCA methods to perform mining. The classification of the reviews of emotions is the core concept of the reviews received by the viewpoint of the patient. The current reviews are displayed by tagging the subject words to Data available in the MEDDRA text form. Manek AS et al [3]. proposed a model which is designed by using 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% [4]. PAMM for mining aspects relating to specified labels or groupings of drug reviews are used in [5]. Gräßer F et al. discusses an approach where the first sentiment analysis is conducted to predict the sentiments concerning overall satisfaction, side effects and effectiveness of user reviews on specific drugs [11]. To meet the challenge of lacking annotated data, further the transferability of trained classification models among domains, i.e. conditions, and data sources are investigated [7]. In our work, we have applied different machine learning algorithms to find the best suitable one on the modified data set with extra parameters.

III. PROPOSED APPROACH

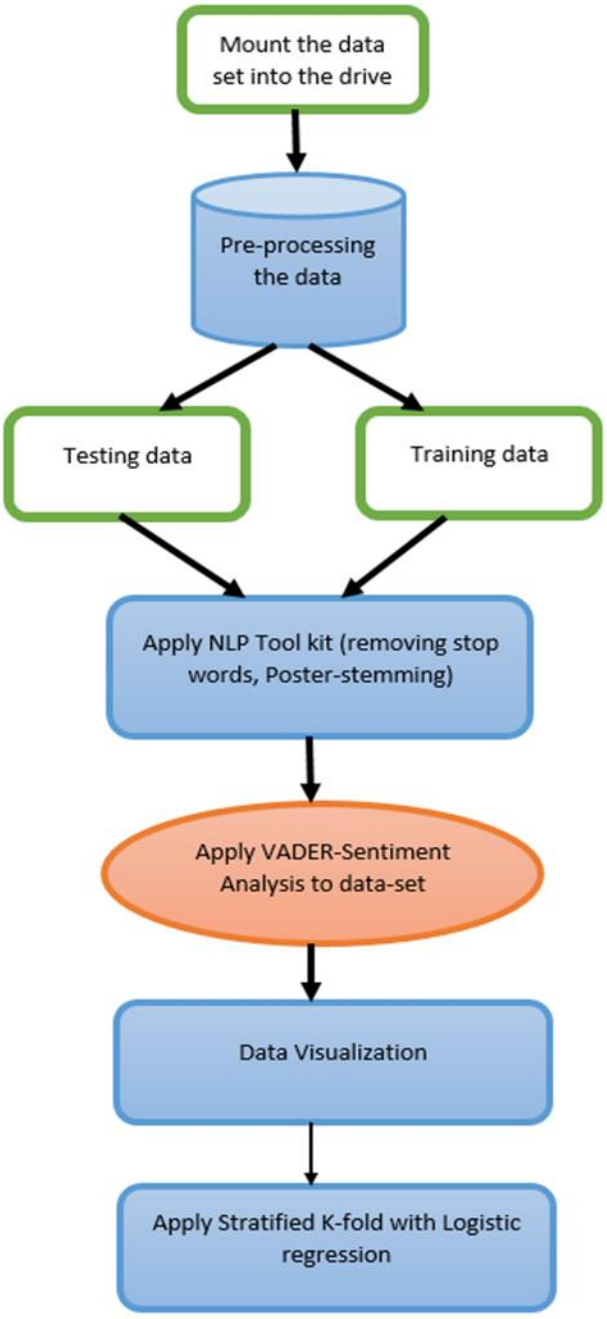

Fig. 1 Proposed Methodology for analyzing sentiment analysis of the data.

A. Data Set Description

We choose "DRUG REVIEW DATA-SET" from the UCI repository where dataset were freely available. Features seen in DRUG REVIW DATA-SET are the name of the drug, condition, review, rating and useful count. The data set consists of 251063 reviews in text format of reviews given by the users. We applied Vader-Sentiment to the data i.e. Pre-processing of the data to extract new features from the data set seen in Table. 1.

B. Implementation

We are using supervised dataset which consists of labelled data. Steps below is the proposed methodology and is shown in Fig 1.

Each review goes to the pre-processing the data by applying Natural Language Processing Removing stop words, Punctuation's, features are generated. Once the raw drug reviews are collected, a number of pre-processing steps are necessary for the generation of clean documents for further processing. These include the following:

1) Tokenize the reviews such that each review is represented as a collection of words for text analysis.
2) Convert all text data to lowercase, so that the words of different cases could be treated the same to remove redundancy.
3) Erase punctuation and symbols, which can safely be ignored without sacrificing the meaning of the sentence.
4) Remove a list of stop words such as 'and' and 'the' that does not add much meaning to a sentence.
5) Lemmatize the words to reduce words to their dictionary forms such that for example, 'am', 'are' and 'is' can all be converted to 'be'.

- After applying Machine Learning techniques

1. Linear Support Vector Machine

The most applicable machine learning algorithm for our problem is Linear SVC., we Visualize the data.

2. Naïve Bayes

Naive Bayes is a simple method for building classification: models for class labels that give problem instances, that are depicted as vectors of functional values, where class labels draw from certain finite set.

3. Logistic Regression

The binary logistic regression version has tiers of the structured variable: categorical outputs with greater than two values are modeled via multinomial logistic regression, and if the more than one classes are ordered, by means of ordinal logistic regression.

4. Support Vector Machine

support-vector machines (SVMs, also support-vector networks[1]) in machine learning are supervised learning models of different learning algorithms modeling data used to analyze classification and regression.

5. Random Forest

Random forest is an ensemble method used for classification of decision trees at training time outputting the data into the label data like A and B.

- Then Stratified K-Fold is used along with logistic regression for performance evaluation.

Table. 1 Drug review Data Set Feature extraction using VADER-Sentiment Analysis

ID	Drug Name	condition	Review	rating	date	Useful Count	Rating Sentiment	Rating Sentiment Label	Vader Review Score
163740	Mirtazapine	Depression	"I've tried a few antidepressants	10	28-Feb-12	22	2	positive	2
206473	Mesalamine	Crohn's Disease, Maintenance	"My son has Crohn's	8	17-May-09	17	2	Positive	2
159672	Bactrim	Urinary Tract Infection	"Quick reduction of symptoms"	9	29-Sep-17	3	2	positive	0

IV. RESULT AND DISCUSSION

The results of the system are the aspects in terms of drug reviews evaluated under Stratified K-fold with Logistic regression approach and interpretation are made out of it. Here in this work drug reviews are evaluated. Where the reviews were classified to positive and negative reviews using Sentiment Analysis. For Clean review data applied VADER-sentiment analysis. We feed the input data to methods as shown in Table 2. Where we can see mostly nominal results as shown in Fig .2. We use stratified K-Fold is a variation of k-fold which returns stratified folds: each set contains approximately the same percentage of samples of each target class as the complete set. To improve the Performance of the sentiment reviews.

Table.2 Performance Percentages On Different Comparison Models

Method	Accuracy	F1	Precision
Linear SVC	90.19	90.10	90.10
Naïve Bayes	79.57	78.48	76.92
Logistic Regression	85.27	84.23	86.43
Support Vector Machine	89.34	89.12	82.78
Random Forest	81.83	81.21	82.56
Logistic Regression + Stratified K-Fold	**96.19**	**95.47**	**96.10**

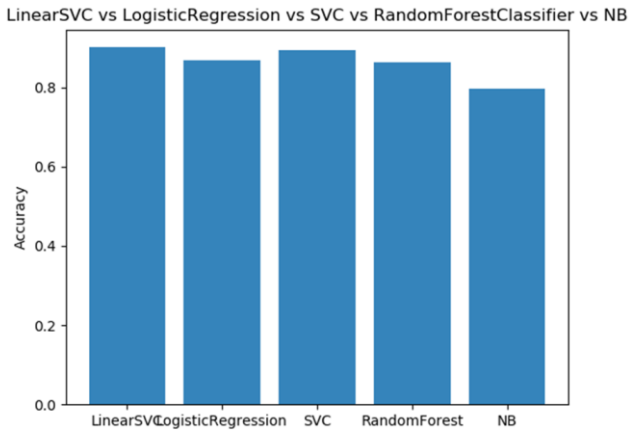

Fig. 2: Visualization for comparison Between models for drug reviews

Combining Logistic Regression with the stratified K-fold for the data selection

To optimize the results of previously discussed algorithms, we have combined stratified K-fold with logistic regression achieve the better accuracy up to 96% as shown in Fig 3:

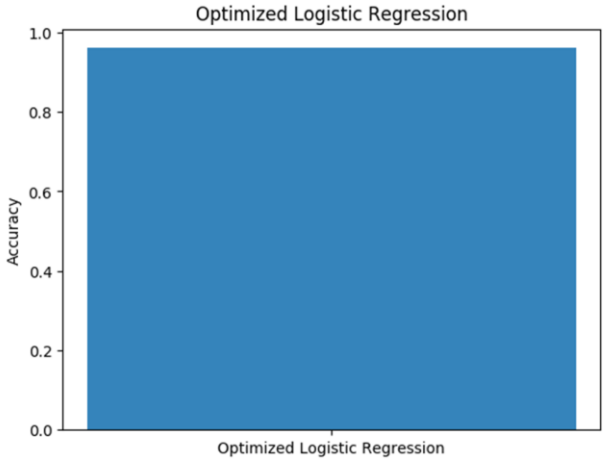

Fig. 3: Effective measure of accuracy on drug reviews using Stratified K-Fold along with Logistic Regression

V. CONCLUSION

This paper proposed an approach to improve the effectiveness of the drug reviews given. Also it added extra parameters to the chosen dataset to improve the effectiveness of the reviews. We have applied VADER sentimental analysis on the "DRUG REVIEW" dataset to get the structured data and then applied the existing machine learning techniques. After careful considerations it has seen that the results were nominal. To achieve the better accuracy, we have applied the stratified algorithm combined with logistic regression. It is observed that the accuracy has increased after applying the stratified approach. These results help pharmaceutical companies to provide a better medication to the patients. As a future work, this work can be extended by applying the stratified algorithm with novel and hybrid machine learning algorithms.

REFERENCES

1. Liu J, Seneff S. A dialogue system for accessing drug reviews. In2011 IEEE Workshop on Automatic Speech Recognition & Understanding 2011 Dec 11 (pp. 324-329). IEEE.
2. Cheng VC, Leung CH, Liu J, Milani A. Probabilistic aspect mining model for drug reviews. IEEE transactions on knowledge and data engineering. 2013 Dec 3;26(8):2002-13.
3. Cheng V, Tang C, Li CH. Drug Review Mining with Regressional Probabilistic Principal Component Analysis. Computer Science Department Hong Kong Baptist University, HI-KDD. 2012;12.
4. Manek AS, Pandey K, Shenoy PD, Mohan MC, Venugopal KR. Classification of drugs reviews using W-LRSVM model. In2015 Annual IEEE India Conference (INDICON) 2015 Dec 17 (pp. 1-6). IEEE.
5. Tekade TN, Emmanuel M. Probabilistic aspect mining approach for interpretation and evaluation of drug reviews. In 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES) 2016 Oct 3 (pp. 1471-1476). IEEE.
6. Gräßer F, Kallumadi S, Malberg H, Zaunseder S. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. InProceedings of the 2018 International Conference on Digital Health 2018 Apr 23 (pp. 121-125).
7. Chawla D, Mohnani D, Sawlani V, Varma S, Khedkar S. Drug review analytics of neurological disorders. In2019 International Conference on Nascent Technologies in Engineering (ICNTE) 2019 Jan 4 (pp. 1-3). IEEE.
8. Min Z. Drugs Reviews Sentiment Analysis using Weakly Supervised Model. In2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA) 2019 Mar 29 (pp. 332-336). IEEE.
9. Kakulapati V, Bhutada S, Reddy SM. Predictive analysis of drug reviews using Gibbs sampling topic modeling. In2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI) 2018 Sep 19 (pp. 2432-2436). IEEE.
10. Zwaida TA, Beauregard Y, Elarroudi K. Comprehensive literature review about drug shortages in the canadian hospital's pharmacy supply chain. In2019 International Conference on Engineering, Science, and Industrial Applications (ICESI) 2019 Aug 22 (pp. 1-5). IEEE.
11. Chen T, Su P, Shang C, Hill R, Zhang H, Shen Q. Sentiment Classification of Drug Reviews Using Fuzzy-rough Feature Selection. In2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) 2019 Jun 23 (pp. 1-6). IEEE.
12. Ru B, Li D, Hu Y, Yao L. Serendipity—A Machine-Learning Application for Mining Serendipitous Drug Usage From Social Media. IEEE Transactions on NanoBioscience. 2019 Apr 4;18(3):324-34.
13. Alimova I, Tutubalina E, Alferova J, Gafiyatullina G. A Machine learning approach to classification of drug reviews in Russian. In2017 Ivannikov ISPRAS Open Conference (ISPRAS) 2017 (pp. 64-69). IEEE.
14. Akiyama T, Sengoku S. The Productivity of Drug Development: A Systematic Review. In2019 Portland International Conference on Management of Engineering and Technology (PICMET) 2019 Aug 25 (pp. 1-13). IEEE.
15. Yadav S, Ekbal A, Saha S, Bhattacharyya P. Medical sentiment analysis using social media: towards building a patient assisted system. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) 2018 May.
16. Chee BW, Berlin R, Schatz B. Predicting adverse drug events from personal health messages. In AMIA Annual Symposium Proceedings 2011 (Vol. 2011, p. 217). American Medical Informatics Association.
17. Lyu G. A Review of Alzheimer's Disease Classification Using Neuropsychological Data and Machine Learning. In2018 11th International Congress on Image and Signal Processing, Bio-Medical Engineering and Informatics (CISP-BMEI) 2018 Oct 13 (pp. 1-5). IEEE.
18. Shastri SS, Nair PC, Gupta D, Nayar RC, Rao R, Ram A. Breast cancer diagnosis and prognosis using machine learning techniques. InThe International Symposium on Intelligent Systems Technologies and Applications 2017 Sep 13 (pp. 327-344). Springer, Cham.
19. Pandey S, Supriya M, Shrivastava A. Data classification using machine learning approach. InThe International Symposium on Intelligent Systems Technologies and Applications 2017 Sep 13 (pp. 112-122). Springer, Cham.

20. . S. V. Chandolu, Dharaa, C., and Dr. Prakash P, "Railway Gate System: Railway Gate Status Detection", International Journal of Recent Technology and Engineering, vol. 7, pp. 523-525, 2019.
21. D. M. Dhanalakshmy, Jeyakumar, G., and Velayutham, C. S., "Crossover-free differential evolution algorithm to study the impact of mutation scale factor parameter", International Journal of Recent Technology and Engineering, vol. 7, pp. 1728-1737, 2019.

AUTHORS PROFILE

D.Naga Swathi, M.tech (Computer Science and engineering) from Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India, B.Tech,,(Computer Science and Engineering) from SREC-Warangal, Her areas of interest are Machine learning and Data Science

Mr.Kumaran.U, Ph.D. Vellore Institue of Technology,Vellore.M.E.,Arunai Engineering College,Tiruvannamalai,Affiliated to Anna University,Chennai.B.E., Arunai Engineering College,Tiruvannamalai,Affiliated to Anna University..He had 10+ years of Experience in Teaching and research. His area of interest includes Machine Learning, Cloud Compting,Data Mining,Network Security,Privacy Preserving Techniques and interest of things.