ISSN 2007-9737

Exploratory Data Analysis and Sentiment Analysis of Drug Reviews

Bijayalaxmi Panda, Chhabi Rani Panigrahi, Bibudhendu Pati

Rama Devi Women's University,
Department of Computer Science,Bhubaneswar,
India

bijayalaxmi.panda81@rediffmail.com,
{panigrahichhabi, patibibudhendu}@gmail.com

Abstract. The exponential increase in the volume of data in our daily life needs to be managed and analyzed properly to get knowledge and benefit out of that. A drug review dataset obtained from the UCI machine learning repository has six parameters namely drug-id, name, condition, review, rating, and usefulness count out of which we have filtered a subset of the dataset based on the eight conditions. Exploratory Data Analysis (EDA) and Sentiment Analysis (SA) are then applied to the filtered data set. EDA shows the total number of medicines used for all light conditions, number of reviews per condition, five most popular drugs based on usefulness count, number of drugs per condition, etc. SA is performed on the filtered dataset, in which twenty-eight drugs are compared based on rating and polarity where three drugs *Lisdexamfetamine*, *Vyvanse*, *Lamotrig ine* are found to be the best in the view of customers as per their rating and positive polarity and *Suvorexant* is the drug found to have negative polarity and least rating.

Keywords. Drug review, exploratory data analysis, adverse drug reaction, subjectivity, polarity.

1 Introduction

In this digital world, people do not have time to spend for their day-to-day activities physically. Therefore, they depend on several electronic activities such as purchasing goods, an appointment with doctors, bank transactions, and so on. In the current time, every need of human beings is satisfied by electronic means. While purchasing goods, a user's first choice is a trusted website then the user checks the customer feedback and rating. By analyzing such things, we get ready for purchasing such products. Many datasets are available related to Amazon product review [19].

Similarly, there are several aspects of health sectors on which we focus such as patient review on choosing a hospital for treatment, health check-up, drug review, a side effect of drugs dosage and effectiveness, etc. [11].

Clinical-social-personality is the standard of measurement in health-related sentiment analysis [12]. This helps the drug makers by giving them opinion of drug users.

The rest part of the paper is organized as follows. Section 2 presents the related work. Section 3 presents the methodology for the proposed approach, which contains data description, exploratory data analysis, and sentiment analysis. Section 4 describes the experimental results and discussion. Section 5 concludes the paper and identifies certain future research directions.

2 Related Work

In the present scenario, the drug is an essential aspect of human life. Many researchers are working on drug reviews so that common people can get an idea about the best drug for a particular disease. Cavalcanti *et al.* [1] suggested a new unsupervised and knowledge-based method for the extraction of aspects in drug reviews.

Hiremath *et* al. [2] focused on a case study to develop a clinical decision support system for

personalized therapy process using aspect- based sentiment analysis.

The process is carried out on drug review data to determine whether the patient's behavior towards a medicine, product, treatment, etc is positive, negative, or neutral using Natural Language Processing techniques. The polarities obtained are compared for further analysis of the patient reviews for a better clinical decision system.

Das *et al*. [3] developed a learning model that can be trained to predict the disease type when provided with a drug name and its corresponding review. To mitigate the above- mentioned issue, the authors presented and compared various machine learning-based prediction models and their performance compared based on metrics such as precision, recall, F1-Score, and accuracy.

Vijayaraghavan et al. [4] worked on analyzing reviews of various drugs which have been reviewed in the form of texts and have also been given a rating on a scale from 1-10. We had obtained this data set from the UCI machine learning repository which had 2 data sets: train and test (split as 75-25%). We had split the number rating for the drug into three classes in general: positive (7-10), negative (1-4), or neutral (4-7). There are multiple reviews for the drugs that belong to a similar condition and we decided to investigate how the reviews for different conditions use different words impact the ratings of the drugs.

Our intention was mainly to implement supervised machine learning classification algorithms that predict the class of the rating using the textual review. We had primarily implemented different embedding such as Term Frequency Inverse Document Frequency (TFIDF) and the Count Vectors (CV). Authors had trained models on the most popular conditions such as "Birth Control", "Depression" and "Pain" within the data set and obtained good results while predicting on the test data sets.

Shiju *et al.* [5] built different classification models to classify user ratings of drugs with their textual review. Multiple supervised machine learning models including Random Forest and Naive Bayesian classifiers were built with drug reviews using TF-IDF features as input. Also, transformer-based neural network models including BERT, BioBERT, RoBERTa, XLNet,ELECTRA, and ALBERT were built for classification using the raw text as input.

Overall, BioBERT model outperformed the other models with an overall accuracy of 87%. Compagner *et al.* [6] focused on characterizing the sentiment of online medication reviews of Selective Serotonin Reuptake Inhibitors (SSRIs) and Serotonin–Norepinephrine Reuptake Inhibitor (SNRIs) used to treat depression. The publicly available data source used was the Drug Review Dataset from the University of California Irvine Machine Learning Repository. This study utilized a sentiment analysis of free- text, online reviews via the sentimentr package.

The result shows that average sentiment was higher in SSRIs compared to SNRIs (0.065 vs. 0.005, $p < 0.001$). The average sentiment was also found to be higher in high-rated reviews than in low-rated reviews (0.169 vs. −0.367, $p < 0.001$). Ratings were similar in the high-rated SSRI group and high-rated SNRI group (9.19 vs. 9.19).

Gräßer *et al.* [7] proposed a new approach that includes different steps. First, extra parameters are added to the review data by applying VADER sentimental analysis to clean the review data. Then, different machine learning algorithms are applied, namely linear SVC, logistic regression, SVM, random forest, and Naive Bayes on the drug review dataset. To improve this, a stratified K-fold algorithm was applied in combination with Logistic regression. With this approach, the accuracy obtained was increased to 96%.

Mishra [19] performed sentiment analysis of the reviews of drugs given by the patients after the usage using the boosting algorithms in machine learning. The dataset used, provides patient reviews on some specific drugs along with the conditions the patient is suffering from and a 10-star patient rating reflecting the patient satisfaction.

EDA is carried out by the customers to get more insight and engineer features. To classify the reviews as positive or negative three classification models such as LightGBM, XGBoost, and CatBoost were trained and the feature importance is plotted.

The results show that LGBM is the best performing Boosting algorithm with an accuracy of 88.89%. In most of the works related to drug review dataset, it was found that researchers used

supervised machine learning algorithms for classification and comparison.

EDA was performed on the whole dataset in [19] to obtain inside features. In this work, we have performed EDA on drugs of specific disease so that relevant drugs can be compared.

For this, we have taken drugs having more number of reviews for comparison.

3 Methodologies

In this section, we have presented the description of the dataset used along with EDA which includes the result analysis of the drug review dataset, and SA which is used to analyze the subjectivity and polarity of the dataset.

3.1 Data Set Description

The drug review data set was collected from the UCI machine learning repository [18]. The dataset contains patient reviews subject to specific drugs, along with conditions and a 10- point rating depending on the fulfillment of the needs of patients. The dataset contains six attributes such as the name of the drug, disease name marked as condition, patient review as review, rating, review entry date, number of users who found review as useful marked as usefulCount.

The data was collected from drug review sites like druglib.com, drugs.com, etc. The data set is represented in two tsv files as train and test. In our work, we have joined the two files and extracted one subset of the data set by filtering out the data based on eight conditions like depression, Insomnia, anxiety, anxiety and stress, Bipolar disorder, major depressive disorder, ADHD, and panic disorder. Then we performed exploratory data analysis on the filtered data set.

3.2 Exploratory Data Analysis (EDA)

EDA is the evaluative process of execution on data to uncover the designs, solve problems, testing hypotheses in virtue of analytical and pictorial representations. EDA is performed based on sample data sets [9]. We can acquire more and more perceptions by performing EDA on sample data set. In this work, we applied certain analytical processes to the filtered data set of drug review data set.

3.3 Sentiment Analysis (SA)

The process of classifying text into positive, negative, and neutral sentiment using different methods of NLP is known as sentiment analysis. This is used in several areas like a movie review, product review, and drug review, and so on. In this research we have focused on drug review, detecting and monitoring adverse drug reactions, or identifying negative or positive sentiment of patients is the part of research where sentiment analysis is applied on UCI drug review data set. [15, 17] We applied text blob analysis on reviews obtained from drug users.

Textblob analysis shows the classification of positive, negative, and neutral reviews. Polarity obtained lies between the range of -1 and +1. Then a comparison of several drugs was made to find out the best one.

4 Experimental Setup and Results

This section describes the experimental setup required for the proposed approach along with results and discussion.

4.1 Experimental Setup

In this work, we have used the drug review dataset for our experimentation. The dataset was taken from the UCI machine learning repository [8]. The dataset contains parameters such as drug name, disease, review of users for a specific drug, rating, etc. EDA and SA are performed on the drug review dataset. We created a subset of the main dataset taking into account eight conditions.

EDA shows an insight of the subset of dataset such as details of medicines used for the said conditions, number of drugs per condition, most popular drugs according to their usefulness. In sentiment analysis part by using textblob in python we performed subjectivity analysis to find out positive and negative opinions of patients who are using these drugs. Polarity was obtained and compared with drugs.

Table 1. Reviews with usefulness count

Id	Condition	Rating	Drug Name	Review	Useful Count
96616	Depression	10	Sertraline	"I remem ber reading people 's opinions, on...	1291
119151	Depression	9	Zoloft	"I&#039 ;ve been on Zoloft 50mg for over two ye...	949
62688	Anxiety and Stress	8	Citalopram	"I work for a large Fire Depart ment. I was ha...	693

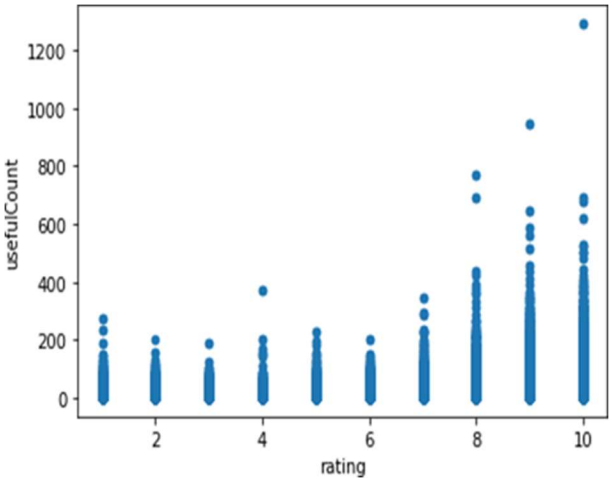

Fig.1. Rating vs. useful count

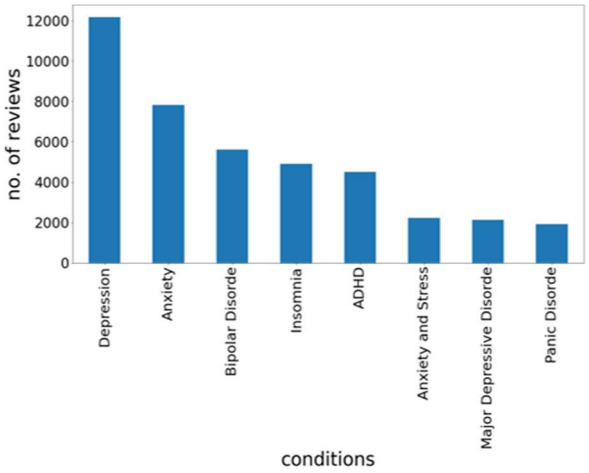

Fig. 2. Number of reviews per condition

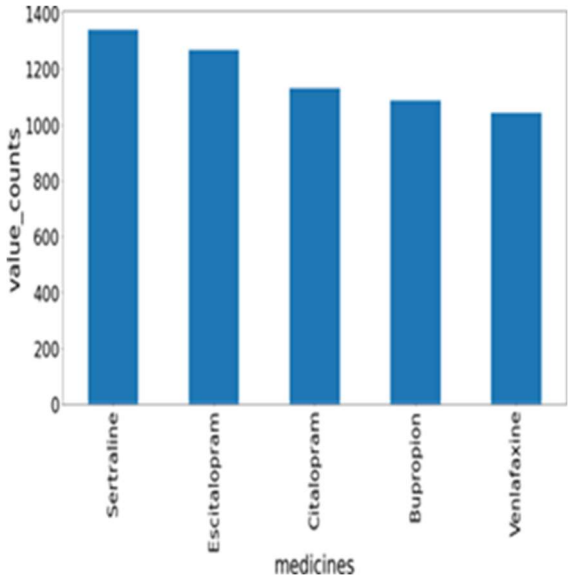

Fig. 3. Popular drugs based on counts

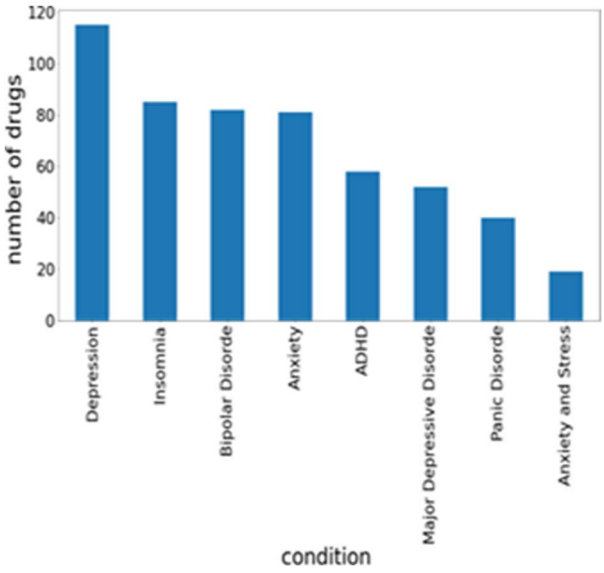

Fig. 4. Number of drugs per condition

Table 2. Mean rating and mean polarity of drugs

Sl. No	Drug name	Condition	Mean rating	Mean polarity
1	Bupropion	depression	7.445	0.097
2	Sertraline	depression	7.497	.066
3	Venlafaxine	depression	6.8	0.053
4	Pristiq	depression	7.218	0.094
5	Desvenlafaxi ne	depression	7.274	0.092
6	Citalopram	depression	7.666	0.081
7	Escitalopram	depression	7.843	0.082
8	Lexapro	depression	7.808	0.081
9	Cymbalta	depression	6.572	0.061

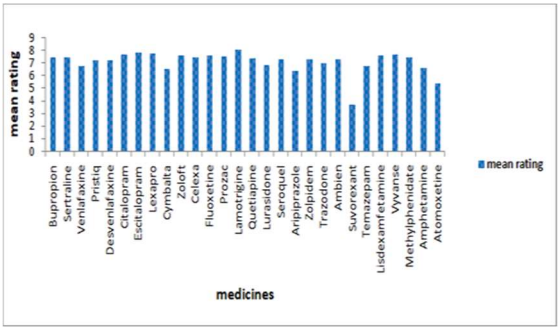

Fig. 5. Mean rating of drugs

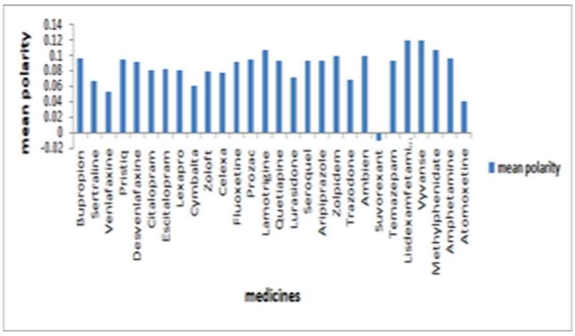

Fig. 6. Mean Polarity of Drugs

Twenty-eight drugs were compared based on their polarity and rating and the best medicine was obtained.

4.2 Results and Discussion

In this section, EDA is performed and results are analyzed.

Exploratory Data Analysis (EDA)

The filtered data set contains six attributes with eight numbers of conditions.

The number of medicines used for those conditions is 299. Table 1 contains the top three reviews on the basis of usefulness count. The users found two most popular drugs useful and are Sertraline and Zoloft. In this case, condition is depression and useful Count are 1291 and 949 respectively.

We drew a scatter plot on rating verses useful Count where the medicine Sertraline and Zoloft which is having 10 rating has highest useful count as 1291.

Some users also have given those medicines 9 rating where useful count is 949 which is clearly understood from the scatter plot as shown in Fig. 1.

The bar graph as shown in Fig. 2 shows the number of reviews per condition. This graph shows highest number of reviews in depression condition and lowest number of reviews in panic disorder condition.

The graph shows top five most popular drugs on the basis of their usefulness count where we found Sertraline, Escitalopram, Citalopram, Bupropion, Venlafaxine are top five most popular drugs and is shown in Fig. 3.

From Fig. 4, it is clear that Depression has highest number of drugs that is 115 and the condition Anxiety and stress has lowest number of drugs such as 19.

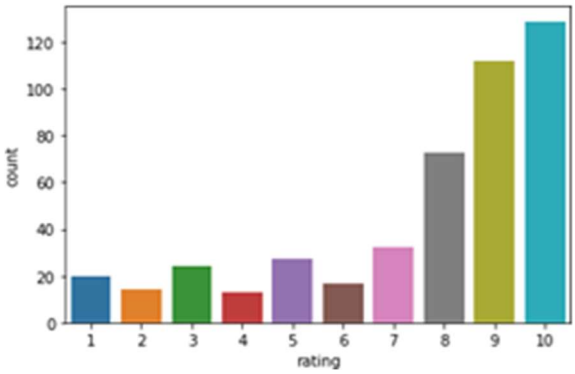

Fig. 7. Rating vs Count of Lisdexamfetamine

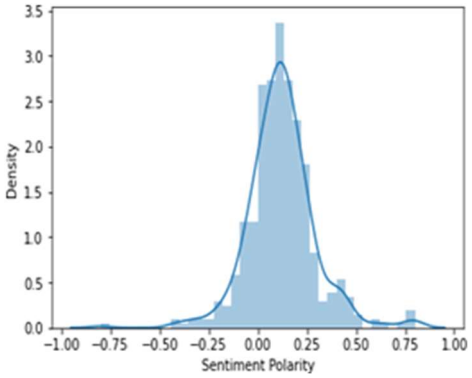

Fig. 8. Sentiment polarity vs. Density of Lisdexamfetamine

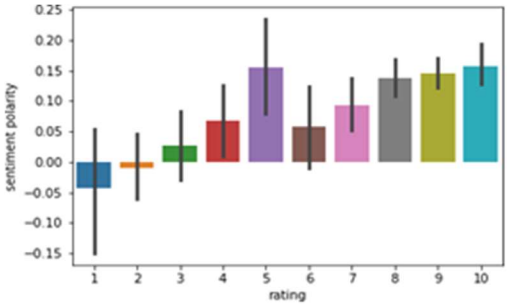

Fig. 9. Rating vs. sentiment polarity of Lisdexamfetamine

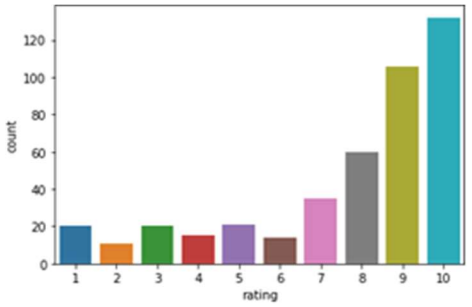

Fig. 10. Rating vs. Count of *Vyvanse*

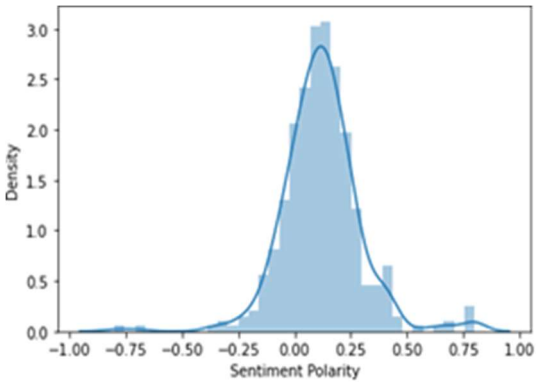

Fig. 11. Sentiment polarity vs. Density of Vyvanse

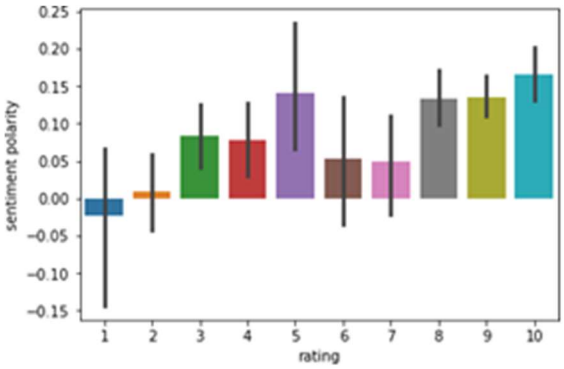

Fig. 12. Rating vs. Sentiment polarity of Vyvanse

Similar kind of drugs may be used for all the conditions. This is shown in Fig. 4.

Sentiment Analysis

Social media data are useful based on healthcare, disease diagnosis and so on. Sentiment analysis is the way to facilitate analysis of information obtained from social media and gives benefit to same kind of users [13]. Self-reported patient data can be obtained from social media that gives positive impact on other patients [14, 16]. On the basis of 8 conditions we have filtered the subset from drug review data set. Many drugs are suggested for each and every conditions. All the

conditions are psychiatric conditions according to which popular drugs are selected and preprocessed the number of reviews [10].

We have performed sentiment analysis using textblob module of python for text classification as positive, negative and neutral.

Then analysis is performed to define polarity numerically, where polarity ranges from -1 to +1. We have taken around 28 medicines on the basis of number of reviews. We have considered those medicines when the number of reviews must exceed 100.All the 28 drugs related to at least 1 condition and at most 8 conditions because one drug can be used to treat several conditions. We have compared the drugs on the basis of their mean rating and mean polarity.

The mean rating of 28 medicines represented in Fig. 5 shows that according to customer rating.

Sentiment analysis according to user reviews obtained from our experimental study is clearly mentioned in Fig. 6. It defines polarity in the range -1 to +1. In other words, it defines positive or negative polarity. According to graphical representation Lisdexamfetamine, Vyvanse, and Lamotrigine are having corresponding polarity values 0.1204, 0.12, and 0.1075 respectively.

Lisdexamfetamin and Vyvanse, both are used to treat the condition Attention deficit hyper activity (ADHD) and Lamotrigine is used to treat the condition Bipolar disorder. From the results it was found that all three medicines are having nearly same positive polarity which is widely accepted by the customers. Suvorexant is the medicine used for treatment of the condition.

Lamotrigine is the best medicine which has mean rating of 8.113 which is mainly used for treatment of bipolar disorder. Polarity of this medicine is 0.107. Customers given lowest rating to the medicine. Suvorexant used to treat Insomnia that is 3.761.So customers are not appreciating this medicine properly, so it needs to be improved.

Insomnia which is having lowest polarity value of -0.0108. This shows negative polarity which is not appreciated by customers and needs to be improved.

The detailed analysis and description of all the three medicines are given below.

The bar graph as shown in Fig. 7 shows rating verses count of Lisdexamfetamine and it shows that maximum customers have given 10 star rating

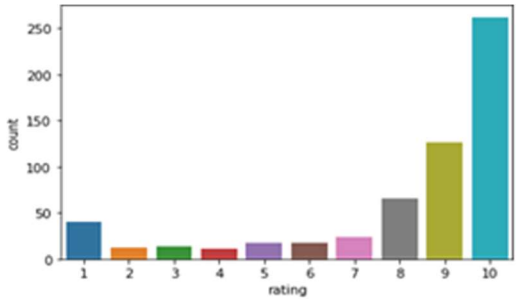

Fig. 13. Rating vs. Count of Lamotrigine

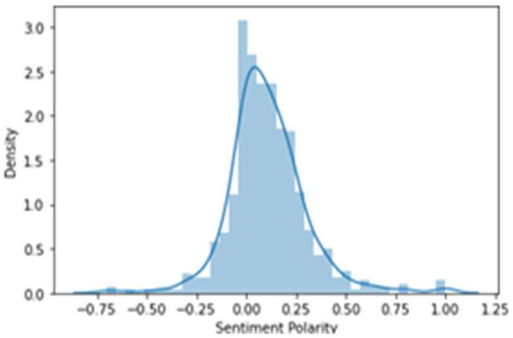

Fig. 14. Sentiment Polarity vs. Density of Lamotrigine

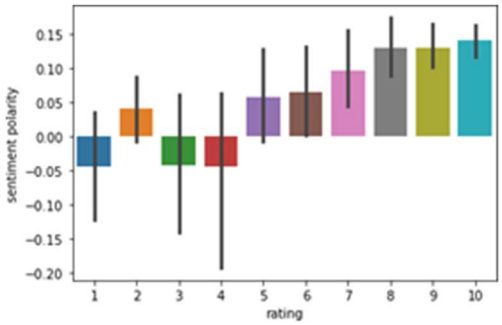

Fig. 15. Rating vs. Sentiment polarity of Lamotrigine

followed by 9, 8, and 7 rating respectively. The histogram as shown in Fig. 8 shows data distribution as sentiment polarity verses density.

The frequency distribution is higher in case of positive polarity and lower in case of negative polarity, so we can assume that users of this medicine have positive sentiment towards it. Fig. 8 shows that the highest frequency lies between 0 and 0.25. Fig. 9 shows rating verses sentiment polarity of Lisdexamfetamine. It is found that most of the customers have given rating 5 where polarity is highest. Many customers have also given 8, 9

and 10 rating where polarity also increases accordingly.

Vyvanse is same kind of medicine as Lisdexamfetamine with negligible difference in number of reviews. Fig. 10, 11, and 12 shows the features of Vyvanse.

The histogram as shown in Fig. 14 shows sentiment polarity verses density where in spite of some existing negative comments there are much more positive reviews of customers who have used this medicine. So in this case positive polarity is found to be higher and hence the product can be considered as reliable.

Lamotrigine sentiment analysis proves it as the 2nd best suggested medicine in our drug review analysis. The graph as shown in Fig. 13 shows rating verses count of Lamotrigine, where users have given highest 10 star rating to this product followed by 9 and 8. So the number of positive reviews is more for Lamotrigine.

From Fig. 15, it is found that many users have given 10 star rating followed by 9, 8 and 7. Sentiment polarity is highest in case of rating 8. It also shows that rating 9 and 10 has more positive polarity.

5 Conclusion and Future Work

In this particular research, EDA and SA is applied on the drug review dataset obtained from UCI machine learning repository. On the basis of eight considered conditions, dataset was compiled because the customers had given more reviews on those conditions.

In EDA, the total number of medicines used for all eight conditions, number of reviews per condition, five most popular drugs on the basis of usefulness count, number of drugs per each condition etc. are described where as in SA, 28 drugs were compared on the basis of rating and polarity.

The drugs such as Lisdexamfetamine, Vyvanse, and Lamotrigine are found to the best drugs in the view of customers as per their rating and positive polarity. Suvorexant was found to be the drug having negative polarity and least rating. Other than eight considered conditions there are also several other conditions and drug classes present in the data set and we can categorize the drugs into groups and can also analyze the drug review dataset for different conditions.

References

1. **Cavalcanti, D. C., Prudêncio, R. B. C. (2017).** Unsupervised aspect term extraction in online drugs reviews. 30th International Florida Artificial Intelligence Research Society Conference (FLAIRS), pp. 38–43.
2. **Hiremath, B. N., Patil, M. M. (2020).** Enhancing optimized personalized therapy in clinical decision support system using natural language processing. Journal of King Saud University-Computer and Information Sciences, Vol. 34, No. 6, pp. 2840–2848. DOI: 10.1016/j.jksuci.2020.03.006.
3. **Das, S., Kumar-Mahata, S., Das, A., Deb, K. (2021).** Disease prediction from drug information using machine learning. American Journal of Electronics & Communication, Vol. 1, No. 4, pp. 16–21. DOI: 10.15864/ajec.1403.
4. **Vijayaraghavan, S., Basu, D. (2020).** Sentiment analysis in drug reviews using supervised machine learning algorithms, arXiv:2003.11643. DOI:10.48550/.arXiv.2003.11643.
5. **Shiju, A., He, Z. (2021).** Classifying drug ratings using user reviews with transformer-based language models. MedRxiv. DOI: 10.1101/2021.04.15.21255573.
6. **Compagner, C., Lester, C., Dorsch, M. (2021).** Sentiment analysis of online reviews for selective serotonin reuptake inhibitors and serotonin–norepinephrine reuptake inhibitors. Pharmacy, Vol. 9, No. 1. DOI: 10.3390/pharmacy9010027.
7. **Gräßer, F., Kallumadi, S., Malberg, H., Zaunseder, S. (2018)**. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. Proceedings of the International Conference on Digital Health (DH), pp. 121–125. DOI: 10.1145/3194658.3194677.
8. **Patil, P. (2021).** What is exploratory analysis? Towards data science (TDS).
9. **Zolnoori, M., Fung, K. W., Patrick, T. B., Fontelo, P., Kharrazi, H., Faiola, A., Shah, N.**

D., Wu, Y. S. S., Eldredge, C. E., Luo, J. (2019). The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. Data in brief, Vol. 24. DOI: 10.1016/j.dib.2019.103838.

10. **Cavalcanti, D., Prudêncio, R. (2017).** Aspect-based opinion mining in drug reviews. In: **Oliveira, E., Gama, J., Vale, Z., Lopes-Cardoso, H., eds.**, Progress in Artificial Intelligence (EPIA), Lecture Notes in Computer Science, Springer Cham, Vol. 10423, pp. 815–827. DOI: 10.1007/978-3-319-65340-2_66.
11. **Cohen, J. (1960).** A coefficient of agreement for nominal scales. Educational and psychological measurement, Vol. 20, No. 1, pp. 37–46. DOI: 10.1177/00131644600 2000104.
12. **Denecke, K. (2015).** Sentiment analysis from medical texts. Health Web Science: Social Media Data for Healthcare, Springer, Cham, pp. 83–98. DOI: 10.1007/978-3-319-20582-3_10.
13. **Gopalakrishnan, V., Ramaswamy, C. (2017).** Patient opinion mining to analyze drugs satisfaction using supervised learning. Journal of applied research and technology, Vol. 15, No. 4, pp. 311–319. DOI: 10.1016/j.jart. 2017.02.005.
14. **Korkontzelos, I., Nikfarjam, A., Shardlow, M., Sarker, A., Ananiadou, S., Gonzalez, G. H. (2016).** Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. Journal of biomedical informatics, Vol. 62, pp. 148–158. DOI: 10.1016/j.jbi.2016.06.007.
15. **Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J., Gonzalez, G. (2010).** Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. Workshop on biomedical natural language processing, Association for Computational Linguistics (BioNLP), pp. 117–125.
16. **Mishra, A., Malviya, A., Aggarwal, S. (2015).** Towards automatic pharmacovigilance: analysing patient reviews and sentiment on oncological drugs. IEEE International Conference on Data Mining Workshop (ICDMW), pp. 1402–1409. DOI: 10.1109/ICDMW.2015.230.
17. **UCI (2021).** UCI Machine Learning Repository. UCI Center for Machine Learning and Intelligent Systems.
18. **Srikanth, K., Murthy, N. V. E. S., Prasad-Reddy, P. V. G. D. (2021).** Sentiment classification on online retailer reviews. 3rd International Conference on Communications and Cyber Physical Engineering, Lecture Notes in Electrical Engineering, Vol. 698, pp. 1557–1563. DOI: 10.1007/978-981-15-7961-5_140.
19. **Mishra, S. (2021).** Drug review sentiment analysis using boosting algorithms. International Journal of Trend in Scientific Research and Development (IJTSRD), Vol. 5, No. 4, pp. 937–941.

Article received on 01/12/2021; accepted on 03/03/2022.
Corresponding author is Chhabi Rani Panigrahi.