Sentiment Prediction of Drug Reviews using Datamining

Gargeyi Baipa, Naveen Donthineni

ABSTRACT:

From the experience of the end user of a drug, which in the context of this project is a patient, we can derive some critical information that might be useful for improving the condition of a patient, can help other patients with similar condition and also help improve the sentiment of certain drug brands. Some studies were performed on how social media influenced the patient's usage of drugs. Drug review data from www.drugs.com, which captured the reviews of the patients, was analyzed in this work. Insights into the usefulness of the drugs and the associated health conditions described were unearthed. Machine learning algorithms were used to predict the sentiment of the reviews. The approach to perform sentiment analysis and the valuable techniques for cleaning the reviews were explored. An assessment of multiple machine learning models was performed to determine which algorithm works best for the chosen dataset.

INTRODUCTION:

There is an enormous amount of data available in the medical industry describing the patient's conditions, the prescribed medicines and the patient's opinion on these drugs. Much of this information is very complex in the form of plain text. Natural Language Processing, also referred to as NLP, helps in understanding the relationship between the words and making an inference about the words. Sentiment analysis, also known as opinion mining, which uses NLP to study the affective state and subjectivity of the text was explored in this project. Supervised machine learning algorithms namely, Logistic Regression, Navies Bayes, Decision Tree, Random Forest and Support Vector Machine were used to predict the sentiment of the reviews. The goal of this project is to propose the model that performed the best for this dataset. The data set was split to utilize 70% for training the models and 30% as test data for evaluation of the models. The entire work was carried out in Python™.

LITERATURE REVIEW:

Vast amount of work was done on performing sentiment analysis on various review data sets, not only in medical industry but also on consumer products, movies etc., Pertaining to the domain of patient reviews on drugs, work was performed to analyze patient satisfaction [1], cross domain evaluations of drugs reviews and evaluation of the polarity of the reviews [2], analysis on side affects and effectiveness of the reviews on specific drugs [3], other work that was followed closely included prediction of sentiment based on the condition [5].There was also a Kaggle University Club Hackathon in winter 2018, in which multiple individuals worked on modeling the sentiment of reviews using several machine learning techniques, prediction of the rating based on review, prediction on the condition of the patient based on the review etc.

DATA REVIEW:

The data used in this project was obtained from UCI machine learning repository. The dataset consists of patient reviews on specific drugs, conditions and over all patient rating on a scale of 1-10. There are a total of 161297 records and 6 attributes. The attributes are, namely, drug name, condition, review, rating, date and useful count, which is the count of number of users who found the review useful. The mean values for drug review and number of users who found review is useful are 6.9 and 28 respectively. This data was obtained by crawling online pharmaceutical website stated www.drugs.com [3]. Fig1 below shows a snapshot of the data set considered.

	Unnamed: 0	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9.0	May 20, 2012	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8.0	April 27, 2010	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5.0	December 14, 2009	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8.0	November 3, 2015	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9.0	November 27, 2016	37

Figure1: Attributes in the raw data

The dataset used was very clean for the most part. The attribute 'condition' had some missing values which contributed to about 0.5% of the total count. Since this percentage of missing value is so small, these values were dropped. The column names were renamed to be in all lower case. There were no duplicate values. Unique values were also investigated. These were the initial cleaning steps of the data before the EDA was performed.

EXPLORATORY DATA ANALYSIS (EDA):

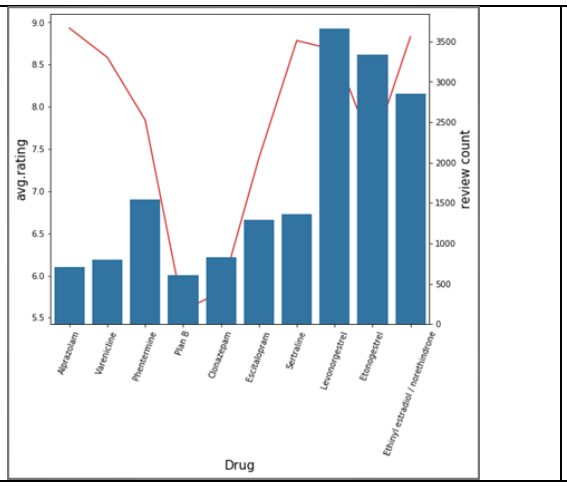
An in-depth Exploratory Data Analysis (EDA) with focus on the list of topics below was performed:

- Distribution of top 10 and bottom 10 ranked drugs, based on ratings and review counts.
- Distribution of top 10 and bottom 10 ranked health conditions by reviews and review counts.
- Changes in drug review counts by year over year for the top 10 drugs which have the greatest number of reviews

Analysis by Data Visualization:

The goal of this section is to visualize and derive key insights of drugs, health conditions and review features present in the dataset.

Analysis of the top-rated drugs and their volume of reviews over the entire time period is shown in Figure 2. The chart shows that drugs with average rating between 5.5 to 8.9 appeared to be the top performing drugs. Drug named Etonogestrel had high review count and low review ratings on the other hand drug named Alprazolam had lowest review count and highest review rating.



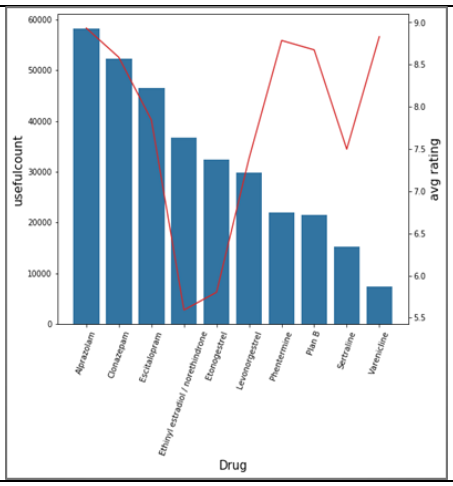


Figure 2: Average rating Vs Review count by Top10 Drugs (left) Average rating Vs Useful count by Top10 Drugs (right)

It was also found that, drugs which had lower review ratings were found to be useful by higher number of users compared to highest average rated drugs. As shown in Figure 2, highest rated drugs like Varenicline had fewer number of users who reviewed the drug and consequently there were a smaller number of users who found drug review useful. At the same time, the reviews for the Etonogestrel were found to be useful by higher number of users even though it had lower average review rating because of large number of reviews that were available.

It was also noted that some of the health conditions like Pain, Depression, Anxiety and Birth Control were having a greater number of drugs in the market – this was examined with two different point of views 1) top 10 health conditions which had most number of reviews 2) number of drugs available for each health condition and distribution of average review ratings of the drugs. As seen in Figure 3 below, Birth Control, Acne and Pain had a greater number of drugs in the market and also the drugs of these health conditions were some of the lowest rated. Most of the weight loss drugs were rated high even though a smaller number of drugs are available in the market when compared to birth control.

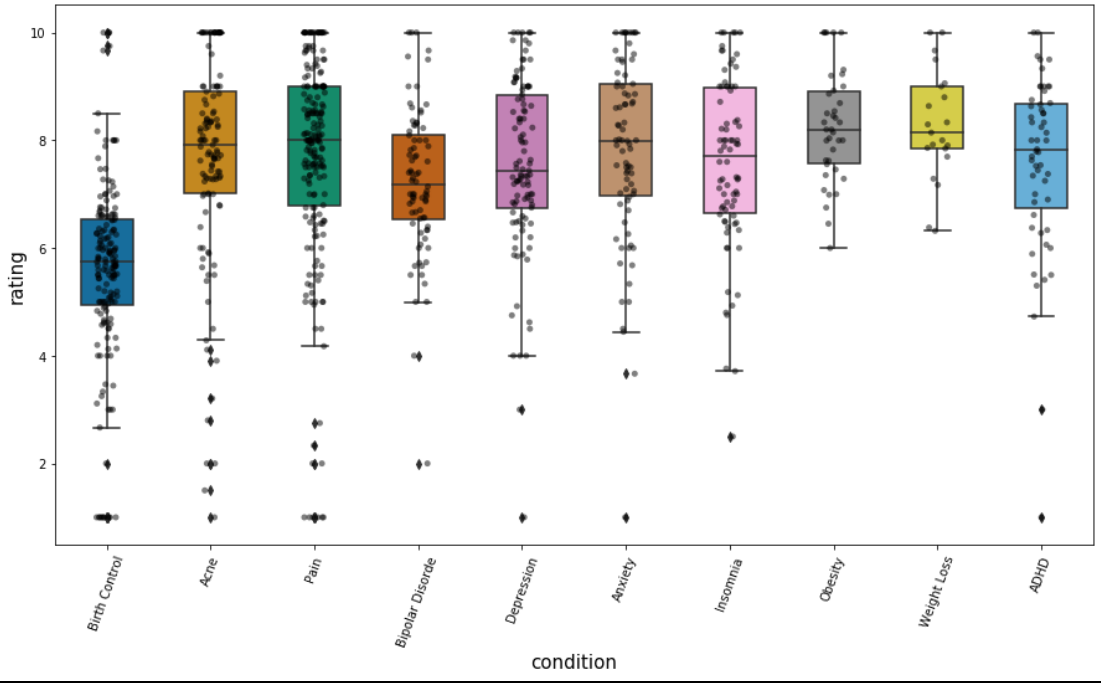


Figure 3: condition vs rating

The next interesting feature is 'date', which helped us to see a trend of the drug review counts over a period. As seen in Figure 4, review counts for all drugs fluctuated from 2008 to 2014 before it peaked to approx. 35000 review counts in 2016.Again there was a sudden dip in the review counts after 2016.On the other hand in Figure 5 number of users found reviews useful increased in 2009 before it dropped to lowest point in 2010 and this trend fluctuated over next 4 years. In 2015, greater number of users found drug reviews were useful before it dropped again in 2016.

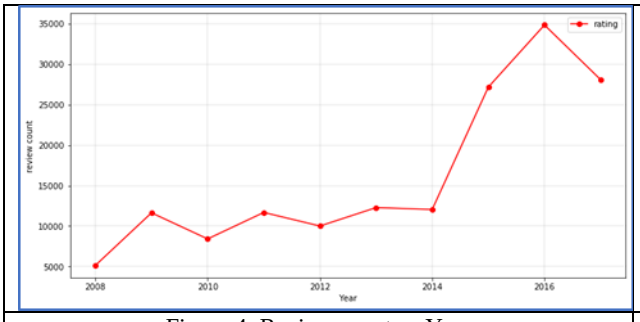


Figure 4: Review count vs Year

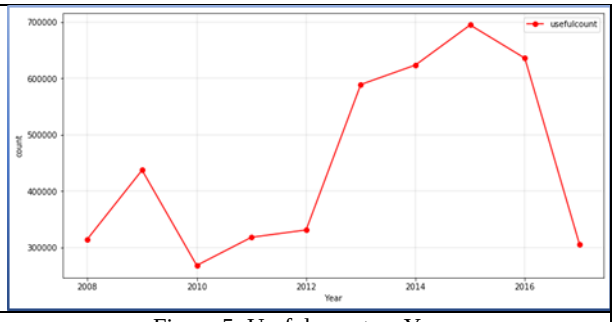


Figure 5: Useful count vs Year

The most interesting feature in the dataset is "Review", which can be analyzed by using Natural Language Processing (NLP) technique. Text analytics was performed on the reviews corresponding to the top 10 drugs and health conditions using NLP in Python™, which is depicted below in Figure 6. The analysis revealed the most frequently occurring words in both the segments is "side effect".

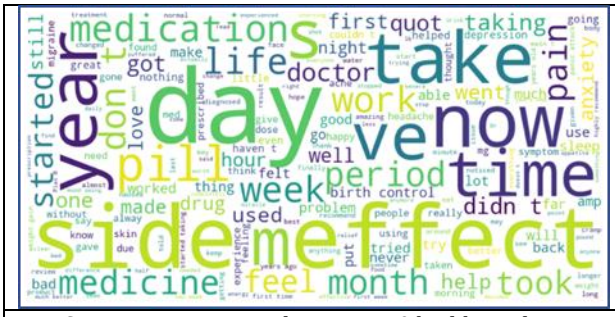


Fig 6: Reviews corresponding to top 10 health conditions

Fig 7: Reviews corresponding to top 10 Drugs

SENTIMENT ANALYSIS:

NLTK package in Python™ was used to clean the attribute review. Cleaning involved multiple steps, the stop words which do not add value to the sentiment are first removed, later punctuations and mistyped words were eliminated. The next approach explored was Stemming of the words. With Stemming, the last few misleading characters were removed. 'PorterStemmer' was used to evaluate this. Considering the limitation of stemming, lemmatization was used in this work. Lemmatization is a process which converts the words into meaningful base word depending on the context, this is referred to as lemmas. In this work, Lemmatization technique was used with the help of Python's 'textblob' library to extract the most important words that can be used for deriving the sentiment.

	drugName	condition	review	rating	usefulCount	wordcount	charcount	lemmatize	sentiment	subjectivity	class_label
53761	Tamoxifen	Breast Cancer, Prevention	"I have taken Tamoxifen for 5 years. Side effe...	10.0	43	97	533	taken tamoxifen year side effect severe sweati...	-0.083333	0.20000	negative
53762	Escitalopram	Anxiety	"I've been taking Lexapro (escitaploprgra...	9.0	11	130	763	taking lexapro escitaploprgram since february ...	0.117193	0.60355	positive
53763	Levonorgestrel	Birth Control	"I'm married, 34 years old and I have no ...	8.0	7	149	780	married 34 year old kid taking pill hassle dec...	-0.049126	0.55676	negative
53764	Tapentadol	Pain	"I was prescribed Nucynta for severe neck/shou...	1.0	20	34	200	prescribed nucynta severe neckshoulder pain ta...	0.000000	0.00000	neutral
53765	Arthrotec	Sciatica	"It works!!!"	9.0	46	2	13	work	0.000000	0.00000	neutral

Figure 8: Processed Review attribute to obtain the sentiment

The lemmatized review was then used for determining the polarity and subjectivity of the review. Polarity is a score that ranges from -1 to 1. Where -1 refers to the negative sentiment associated with a group of words and 1 refers to the positive sentiment associated. The polarity is renamed as sentiment as shown in Figure 8. A new class label is then created to differentiate the sentiment into 3 labels, namely, positive, negative and neutral.

Subjectivity refers to the subjective nature of the sentence, this ranges from 0 to 1. If a subjectivity of 0.9 is observed, it can be inferred that the review is highly subjective based on subjective opinion and not factual.

Analysis showed that as the sentiment increased, the value of the rating also increased as shown in Figure 9. Suggesting a positive sentiment is associated with higher rating. Further, the

distribution of classes showed that the positive sentiment counted to 62%, negative accounted to 33% and neutral sentiment accounted to 5 % of the data which can be seen in Figure 10.

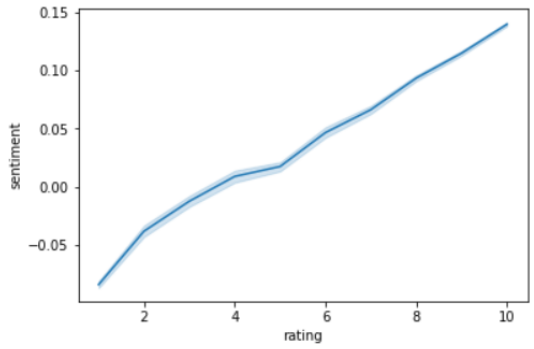


Figure 9: Rating Vs Sentiment

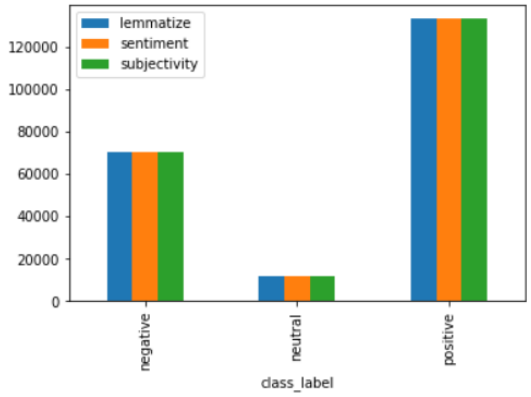


Figure 10: Sentiments distribution in the training set

TERM FREQUENCY AND INVERSE DOCUMENT FREQUENCY:

Before the sentiments that were evaluated can be used for predictive modeling, special preparation is needed. The text must be parsed to split into words, this is referred to as Tokenization. The words are assigned or encoded with unique integers or floating-point values which can be used as an input to ML algorithms. This was achieved using 'scikit' library which helps with the feature extraction. There are multiple options to convert text to the required input. The method that was used in this work is 'TfidfVectorizer'. This summarizes the term frequency and down scales that words that appears a lot.

IMPORTANCE OF FEATURES:

In the work shown thus far, the features that will be used in the model are numerical representation of the words from the attribute 'review'. It would be interesting to find what the most important features or words that will be used in the classification are. A random forest classifier was built and the 'featureImportances' function was used to extract the important features in the analysis. There were 85951 important features in this model. Once the importance of the features was achieved, the values were sorted in descending order to have the most important features on the top shown in Figure 11. Using the vectorizer that created the numerical features the words were then extracted based on the number. The top 6 important features were discovered to be 'bad', 'horrible', 'great', 'worst', 'good', 'terrible' as shown in Figure 12.

	importance
16449	0.014443
39901	0.011668
37404	0.011634
84651	0.010440
37083	0.008879
76637	0.008822

Figure 11: Importance of the feature

	wrd
16449	bad
39901	horrible
37404	great
84651	worst
37083	good
76637	terrible

Figure 12: Feature based on numerical representation

MACHINE LEARNING APPROACH TO PREDICT SENTIMENT OF REVIEWS:

The next step in this work was to use ML algorithms to predict the sentiment of the reviews. The training and test data in the dataset were combined to one data frame in Python. Sampling is performed by splitting the data into Test and training data sets with a 30 % and 70% split. 5 classifiers namely Logistic regression, Decision Trees, Support Vector Machine, Naive Bayes, Random Forest classifiers were used to learn the data.

Key Settings in the model:

Python's 'Scikit-Learn' library was used for the implementing the ML algorithms. In the logistic regression model, since this is a multi-class problem, the multinomial setting was used; this employs the cross- entropy scheme. The class_weight is the next critical setting, the default is all the classes are given the same weight, the balanced mode was used in this work. Balanced mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the given data.

The next model is the Naïve Bayes, since this is text classification with new lot of features, the multinomial Naive Bayes setting was used in this work

The decision tree classifier takes the attributes X and the response Y as input from the training set. This can also handle multi- class problems.

After the results of the decision tree model were witnessed an ensemble classifier, Random Forest, was explored with the expectation that the performance would be improved. The number of trees in the forest was set to 100, the nodes could expand until pure nodes were achieved.

The last model is the Support Vector Machines. In this the linear kernel was used with the expectation of reducing over fitting of the model.

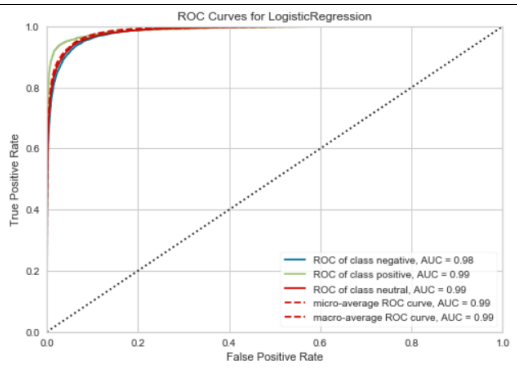
Initially the models were built completely using the 'Scikit' library and its functions. However, there was some difficulty generating good ROC curves. Even though the precision was high, the area under curve was not aligning with the classification report. Upon further research, yellowbrick, a sensical API like 'Scikit' learn was discovered [4]. This helped with obtaining tremendous visualizations that will be explained in the next section.

RESULT DISCUSSION:

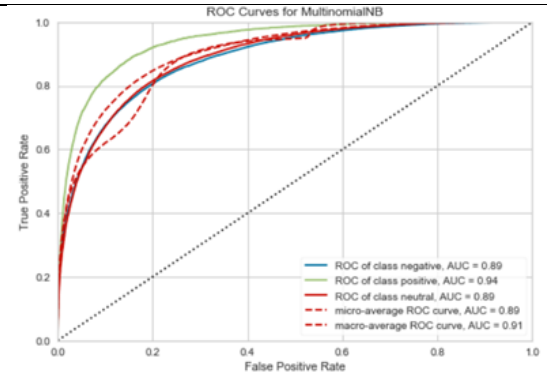
The summary table below shows the accuracy of each of the model. Considering the imbalance in the data as shown in Figure 10, accuracy cannot be used as a measure to judge the model performance. Hence further investigation on the ROC curve of each model was performed.

Model	Accuracy
Logistic Regression	91.8%
Naïve Bayes	67.23%
Decision Tree	90.2%
Random Forest	89.5%
Support Vector Machine	95%

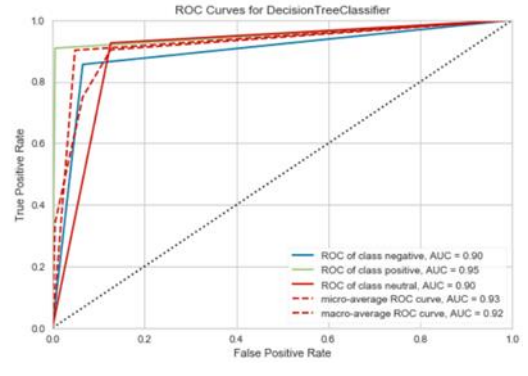
The summary of the ROC curves for each classifier is shown in the Figure 13. It was found that the AUC was least for Naive Bayes classifier. Although the accuracy of Random Forest classifier is lower than decision tree, the AUC is better for Random Forest classifier when compared to decision tree. Other key thing to notice is that even though the number of data points for the neutral review is very low the models showed good AUC value for this label. The model with the best AUC was found to be support Vector Machine with AUC of 0.99 for all the class labels.



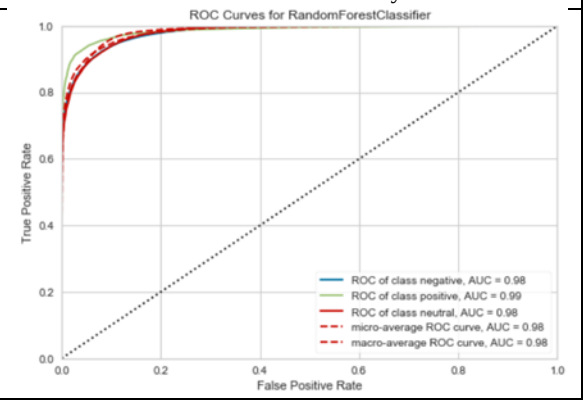
ROC curve for Logistic Regression



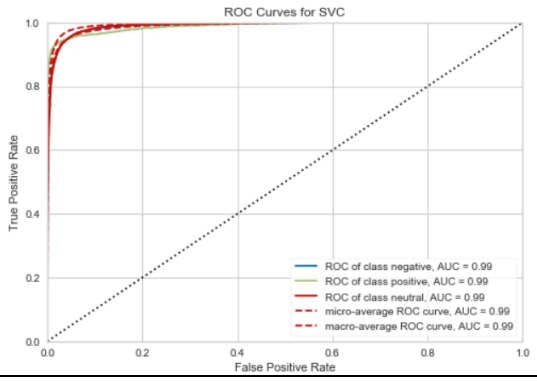
ROC curve for Naïve Bayes



ROC curve for Decision Tree



ROC curve for Random Forest

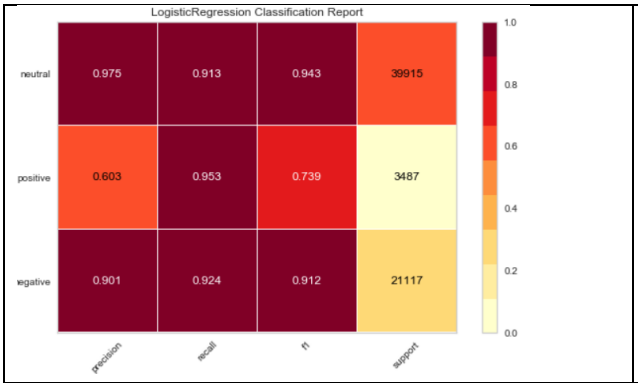


ROC curve for Support Vector Machines

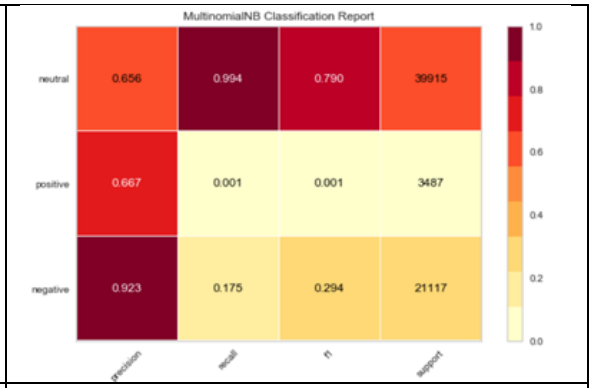
Figure 13: ROC curve summary

It can also be observed that the accuracy of Naïve Bayes classifier is 67%, however, the AUC is better. Looking at the confusion matrix attached in the code, it was understood that this is achieved by the low true negative values.

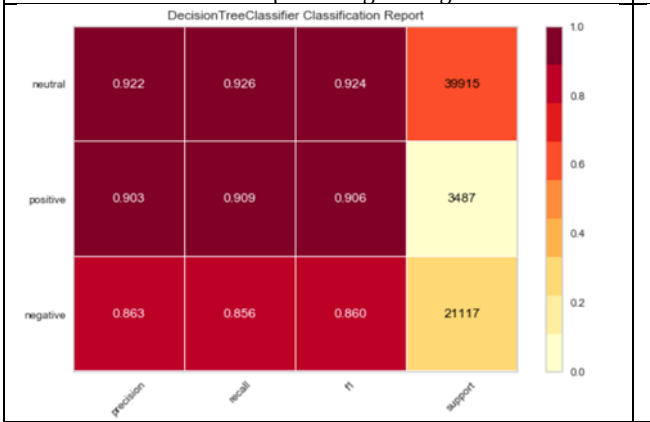
Further investigation into other metrics like precision, recall, F1 were also performed which can be seen in Figure 14. It was found that precision and recall are best for SVM model.



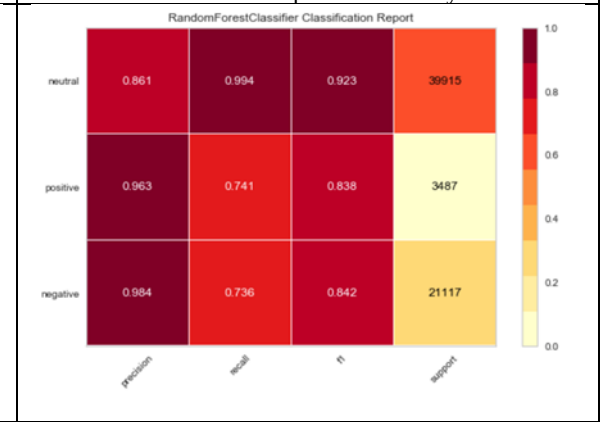
Classification report – Logistic regression



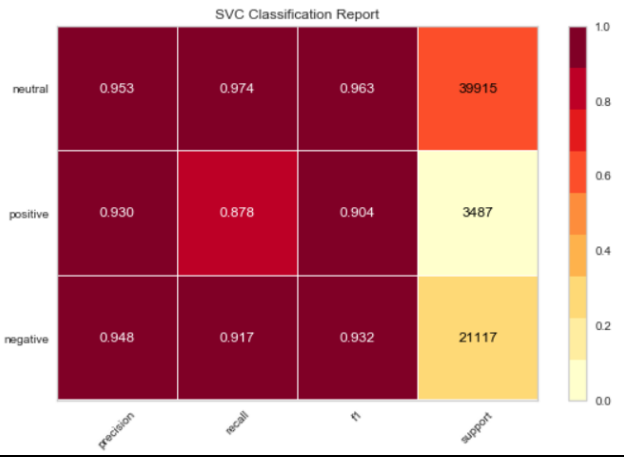
Classification report – Naive Bayes



Classification report – Decision Tree classifier



Classification report – Random Forest



Classification report – Support Vector Machine

Figure 14 : Classification Report Summary

Comparison of results with existing work [5] on this dataset showed some key differences. Work in [5] used neural networks, SVM, Random Forest and Logistic regression and found out that Neural networks is the best model compared to this work where SVM was found to be favorable. Other important difference is, the prediction of the sentiment was made based on a health condition in the cited work [5]. Whereas in this project it was shown that SVM had best performance for predicting the sentiment of a review considering all the conditions. This work employs the use of Tf-idf method opposed to the Count Vectorizer which was used in the existing work for extracting the features from the reviews. Last difference noticed was the use of cross validation in this project sampling by splitting the data into 30 % for test was found to show better performance.

CONCLUSIONS:

In conclusion, the detailed Exploratory Data Analysis revealed some interesting insights into the data. It was found that reviews with most negativity had high useful count. Birth Control and Depression are the top two health conditions and the greatest number of drugs were available for Birth Control and Acne. Further, detailed cleaning was carried on the ‘review’ attribute which is primarily text and the sentiment analysis was performed. The sentiment increased with increase in rating which is a validation of the work done in this step. Further, machine learning models were used to perform the classification to determine the sentiment of the review. The best performing model was found to be Support Vector Machine. Considering the high computation time, it is recommended that Random Forest classifier be used as a second option. Insight into the most important features that was used in the classification model was discovered. For future, this work can be evaluated with Linear SVC function to improve the computation time. Comparison between count vectorizer and tf-idf vectorizer can be performed. Also, it would be interesting to review the performance based on n-gram level features.

References:

[1] https://www.sciencedirect.com/science/article/pii/S1665642317300561

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7233089/pdf/3269957.pdf

[3] https://dl.acm.org/doi/10.1145/3194658.3194677

[4] Yellowbrick Documentation (readthedocs.org)

[5] Sentiment Analysis in Drug Reviews using Supervised Machine Learning Algorithms (arxiv.org)