Prescribing Drugs Using Consumer Reviews*

Mohamed Gomaa[1], **Amar Mohamed**[2]

[1,2] Master of Data Science , FGSSR, Cairo University , Egypt

Abstract

Abstract Social web contains a large amount of information with user sentiment and opinions across different fields. For example, drugs.com provides users' textual review and numeric ratings of drugs.Data mining and machine learning are nowadays among the most known topics in research and they are used to analyze several probabilities of the characterization of databases.we are going to provide a sophisticated and useful model using different NLP or machine learning techniques, using the dataset that has been provided to Find out the most useful drugs for each condition and also try to find out some of the hidden trends and patterns that could help the company to make precise, data driven decisions.

Keywords: *Keyword 1; keyword 2; keyword 3; keyword 4; keyword 5*

1. Introduction

Analysis of drug benefits and side effects plays an important role in determining market trends for consumer use of prescriptions. However, this study may be limited to registered individuals and is often limited to survey participants who met the criteria, making the study more widely available. Geographical scope, there may be exaggerations in the evaluation or classification of the drug and its side effects. To address this issue, the methods of analysis of emotions through social networking sites, as well as analysis of natural language, have been relied upon. There are two main types of drug control after marketing. Some are formed by government regulators such as the Vaccine Adverse Event Reporting System (VAERS) by the FDA [1] or the Yellow Card Scheme by the UK Medicines and Healthcare Products Regulatory Agency [2]. Public/private organizations also have a system for monitoring drug side effects such as searching for harmful drug events and reports and various websites [3]. Past methodologies consist of using statistical measures to identify adverse events such as side effects. The model methodology included consideration of enquiry records or other definitive variables to identify such events but did not focus specifically on the user's feeling through actual reviews [4]. In the present era, there are multiple sources to extract such feelings, the most important of which are social networking sites, as well as pharmaceutical companies' apps and websites, but these sites or apps can be biased, so we focused our solution on linking and detecting the relationships between reviews and evaluations and what the user wrote on the websites. The application of natural language processing offers a different approach to classifying and analysing text information through natural language processing in this project.

2. Methods

2.1. Dataset Preparation

We obtained the dataset from the UCI Machine Learning Repository [17]. These instances were collected from Drugs.com using Beautiful Soup. As can be seen in Table 1, unique I.D. which will be used for identifying all the records in the detail uniquely, then the next one is the drug name that consists of the name of the drugs on which the consumers have given certain ratings and reviews. After that, we have a condition which tells us about the medical condition for which the given drug is used. Next, we have review and rating, which is collected by the patients after they have used that particular drug for their medical condition.After that, we have the date column that tells us that on which date the review and rating was collected by the patient.And finally, we have the useful count, which gives us the information about the number of people who found a drug useful for the specified medical condition.This column is the most important column as this gives a lot of confidence to the people for trusting on a drug for any specified medical condition.Now that we know about the content of the dataset, let us also know about the unique number of drugs and conditions listed in this dataset and the start ending date of collecting these consumer reviews and ratings.

	uniqueID	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	20-May-12	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	27-Apr-10	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	14-Dec-09	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	3-Nov-15	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	27-Nov-16	37

Table 1. The shape of the Dataset

You can see that we have got the statistical summaries for the rating and useful count columns.So let analyze the rating column.We can see that the minimum rating is one, whereas the maximum rating is 10, which is quite obvious,the median is around 7, which is very good.Standard deviation is huge.25% is five, 50% is it, whereas the 75%is ten. Now this means most of the reviews are positive, which is a very good sign. Coming to the useful couch column.We can see that's the minimum count, a zero maximum count is 1291.000000, the average useful count is 28.004755.The standard deviation is 36.403742.Again, this number is very huge.25% is 6.000000, 50% is 16.000000.And the 75% is 36.000000, which means the maximum count is very high compared to other statistical quantities.

	rating	usefulCount
count	161297.000000	161297.000000
mean	6.994377	28.004755
std	3.272329	36.403742
min	1.000000	0.000000
25%	5.000000	6.000000
50%	8.000000	16.000000
75%	10.000000	36.000000
max	10.000000	1291.000000

Table 2. summarize Dataset

After that, we can see that there are no drugs with ratings equal to or greater than eight with zero useful count. Analysis on Useless Drugs The Number of Drugs with No Useful Count : 6318

Number of Good Drugs with Lesser Useful Count : 0 Average Rating of Drugs with No Useful Count : 5.80

Analysis on Useful Drugs The Number of Drugs with Greater than 1000 Useful Counts : 4 Average Rating of Drugs with 1000+ Useful Counts : 10.0

2.2. Model Construction

2.2.1Transformer-Based Models The common methodology for transfer learning has been through the application of pre-training on a large unannotated corpus that was capable of understanding the composition of the data type such as patterns in the language. This process could be considered as self-supervised learning. This pre-trained model is then followed by the fine-tuning process which focused on the training on an application-specific dataset Some common language models are pre-trained by predicting the next word in a sequence, but Bidirectional encoder representation from transformer (BERT) looked at bidirectional predicting context masked intermediate text tokens in the pretraining from Wikipedia and BookCorpus and next sentence prediction. Bert-base-uncased was used for this project.In addition, BERT models such as clinical BioBERT have been pre-trained with a medical corpus from publicly available data from Pubmed and PMC [21]. The model which was used was from Huggingface labeled Bio ClinicalBERT.

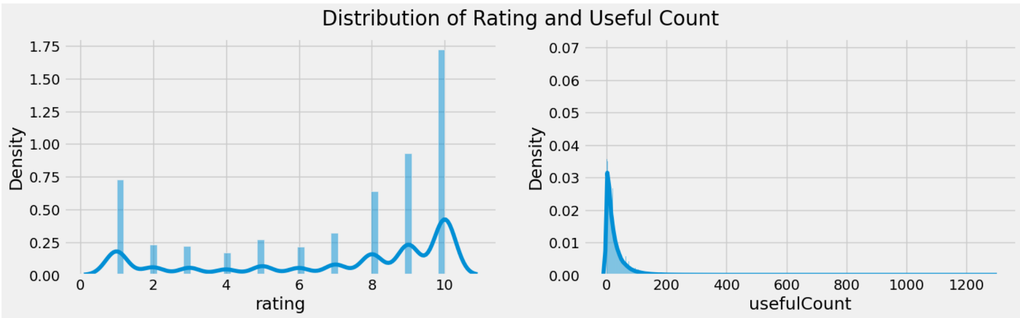

Table 3

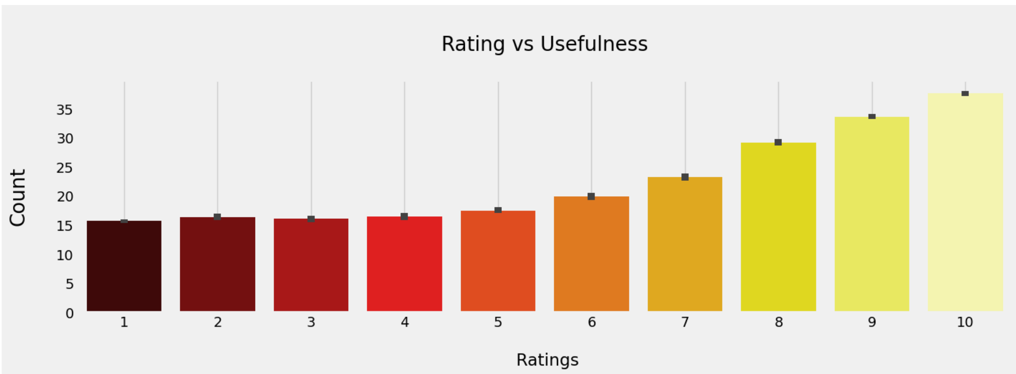

Table 4

2.2.2Classification Model We split the dataset into a training set (60%), a validation set (20%), and the test set (20%). These datasets were further classified into lists which were then converted into Transformer dataset that could be trained by a neural network to generate a model.

2.3. Results

calculate the Sentiment from Reviews Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics and biometrics to systematically identify, extract, quantify and study effective states and subjective information.Now, using sentimental analysis, we will be able to understand the sentimental meaning of each and every review present in the data set now to perform sentimental analysis for this project, we are going to use the Vidra lexicon, which is streamed on millions and millions of decks.After downloading the waiter lexicon model, we will import the sentiment intensity analyzer, and after that we will train all our data and calculate the sentiment score.The sentiment score can range from minus one to plus one, minus one to minus zero point five.Score can be termed as a negative sentiment. Score of minus zero point five to zero point five can be termed as neutral. And finally, between zero point five and one, it can be termed as positive after the sentiments have been calculated.To do that, we are going to use the group by function along with the minimum, average and maximum aggregations. Now taking a closer look at the table where the minimum, average and maximum values of the sentimentshave been listed for each of the ratings, we can see that once again, there is no proper pattern. All of the rating starting from one to 10, have a minimum rating of around minus 0.99. Similarly, the average and maximum ratings for all the categories are also the same, which clearly means that these sentiment scores are of no use at all.

	sentiment		
	min	mean	max
rating			
1	-0.9931	0.005311	0.9898
2	-0.9929	0.003867	0.9924
3	-0.9925	0.003170	0.9877
4	-0.9919	0.000697	0.9867
5	-0.9920	0.014445	0.9882
6	-0.9914	0.008838	0.9936
7	-0.9938	-0.000509	0.9911
8	-0.9936	0.008952	0.9923
9	-0.9964	0.009489	0.9911
10	-0.9982	0.005446	0.9923

Table 5. Impact of Sentiment on Reviews

Calculating Effectiveness and Usefulness of Drugs calculate the effective score for which we are using a formula that we think will work intuitively. deducting the minimum rating from all the ratings and then dividing the rating with maximum rating minus one. After that, we are multiplying the reading with five, which can be treated as the threshold. And after performing the calculations, if the scores come out to be zero one or two, then the effective score would be zero Otherwise it would be one. After the function is created, we can apply this function to create a new column called Effective Score. Now let us calculate the usefulness score. Also to calculate the useful score, we will simply multiply the rating, the useful count and the effective score columns. We can see that most of the useful drugs belong to the depression category, birth control and weight loss, medical conditions, you might cross-check whether the list of drugs are really useful or not by checking the Google reviews or by taking the doctor's recommendation.

	drugName	condition	usefulness
0	Sertraline	Depression	12910
1	Zoloft	Depression	12910
2	Levonorgestrel	Birth Control	12470
3	Mirena	Birth Control	12470
4	Zoloft	Depression	8541
5	Phentermine	Weight Loss	7960
6	Adipex-P	Weight Loss	7960
7	Implanon	Birth Control	7300
8	Viibryd	Depression	6930
9	Vilazodone	Depression	6930

Figure 1. calculate Usefulness Score

Analyzing the Medical Condtitions using the effective scope that we just calculated number of useful and number of useless drugs from each condition. the number of useful drugs for birth control is around 60000, whereas the number of useless drugs for birth control is around 12000. Similarly, if we change the value of condition and select depression from the given list, the number of useful drugs is around 5000, whereas the number of useless drugs is around 2000. check the same results using the bar plot also so that we can understand the overall pattern. But we cannot plot all the conditions as there are more than 800 medical conditions, so we are only going to plot the popular medical conditions like birth control, depression being anxiety, acne, bipolar disorder, insomnia, weight loss, obesity, ADHD, diabetes, type two emergency contraception, high blood pressure, migraine.

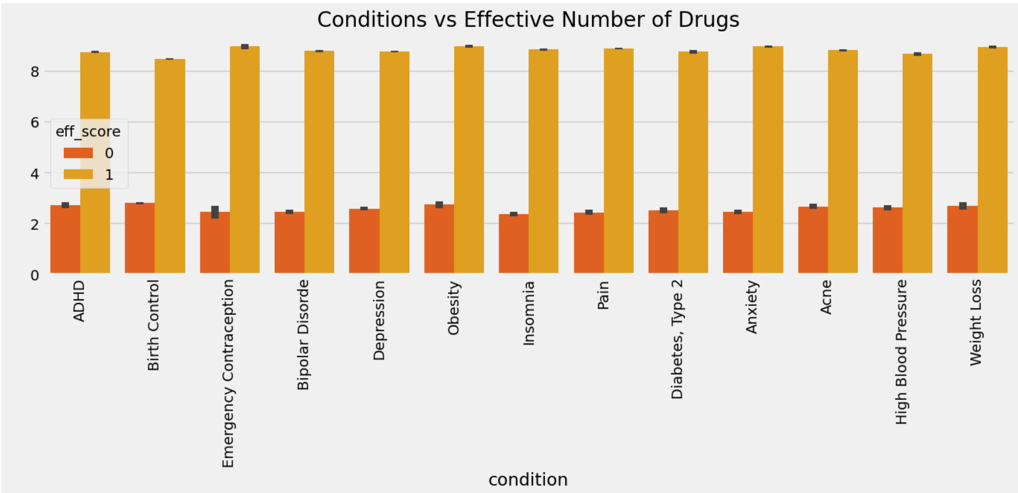

Figure 2. Number of Useless and Useful Drugs for Each Condition

2.4. Conclusion

This study covered the construction of a transformer-based model for the classification of drug reviews , this project can be utilized by doctors, physicians, psychiatrists, physiologists, etc., but not people like you or me, not even data scientist or data analysts. we have completed this project, but you can also try converting this project into a predictive model by training a machine learning model to predict the usefulness of a drug by looking at the consumers review. For that, you will need to learn the LP concepts like Bag of Words, D.F. ideas, etc. in more detail.

References

[1] This study was partially supported by the National Institute on Aging (NIA) of the National Institutes of Health (NIH) under Award Number R21AG061431; and in part by Florida State University-University of Florida Clinical and Translational Science Award funded by National Center for Advancing Translational Sciences under Award Number UL1TR001427. The first author would like to thank eHealth Lab at FSU and the Undergraduate Research Opportunity Program at Florida State University for the mentorship and guidance.

[2] B. O' Donovan, Ruth M. Rodgers, Anthony R. Cox, and Janet Krska. 2019. Making medicines safer: analysis of patient reports to the UK's Yellow Card Scheme. Expert Opinion on Drug Safety 18, 12 (December 2019), 1237–1243. DOI: https://doi.org/10.1080/14740338.2019.1669559Google Scholar

[3] L. Elad Yom-Tov and Evgeniy Gabrilovich. 2013. Postmarket Drug Surveillance Without Trial Costs: Discovery of Adverse Drug Reactions Through Large-Scale Analysis of Web Search Queries. J Med Internet Res 15, 6 (June 2013), e124. DOI: https://doi.org/10.2196/jmir.2614Google Scholar