Analyzing sentiment in drug reviews

1 Introduction

Pharmaceutical companies develop drugs based on intensive research and clinical trials. However, clinical trials provide a limited resource to examine the drugs and their side effects. The patients' reviews after mass production are a great source of information for pharmaceutical companies to learn more about customer reaction to their products, discover potential problems before they occur, and better allocate their resources.

In this project, we will aim to predict whether or not a customer review is negative (rating <4) based on the review, analyze what indicators in a review make it most likely to be negative for different types of drugs and which patients tend to have more negative reviews.

This data set[1][2] contains patients' review and rating of 3436 different drugs. The condition of the patients for which the drug is taken is given and the number of people who found the review useful. There are 161,297 reviews in this data set.

2 Data Wrangling

2.1 Missing data

For some patients the "condition" is missing, and it is replaced by the statement that how many people found that review useful. We replace them all by "unknown" using a regular expression.

2.2 Polarity and subjectivity

Polarity and subjectivity for each review is found using TextBlob, and were added as two separate columns.

---

[1] *Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In Proceedings of the 2018 International Conference on Digital Health (DH '18). ACM, New York, NY, USA, 121-125.*

[2] UCI Machine Learning repository

2.3 Low-rating reviews

In this project we are interested in investigating which patients tend to give low ratings (rating<4) and what elements in the reviews make it most likely to be a low rating. Therefore, we add a binary column and classify the reviews to low rating and high rating groups. Since we are interested in predicting low-rating reviews we will assign 1 to low ratings (ratings 1-3) and 0 to others.

The countplot of the rating column, Fig 2.1, shows that there are almost twice ratings of 10, compared to the second highest ratings which is 9. The data is imbalanced with only 21.6% of low rating reviews.

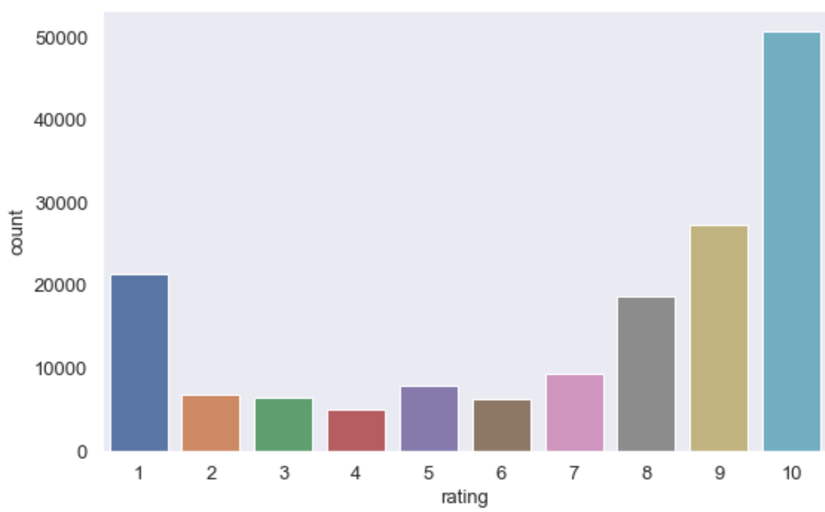

Fig 2.1
Rating countplot

2.4 Drugs and conditions with reviews >100

In order to get a more reliable result, we restrict our data set to drugs and conditions that have at least 100 reviews. This restriction reduces our final data set to 116327 rows.
The top conditions with the highest number of reviews are birth control, depression and pain. The top drugs with the highest number of reviews are Levonorgestrel, Etonogestrel, and Ethinyl estradiol / norethindrone. The first two drugs are used for birth control and the third one is used to treat symptoms of menopause.

Top conditions	Number of reviews	Top drugs	Number of reviews
Birth Control	3631	Levonorgestrel	3631
Depression	3321	Etonogestrel	3321
Pain	2750	Ethinyl estradiol / norethindron	2750

Fig 2.2
Conditions and drugs with the highest number of reviews.

3 EDA

3.1 Top low-rated drugs

The top low rated drug is ‘Miconazole’ with 70% low rating reviews. ‘Miconazole’ is used to treat vaginal yeast infection. The main complaint about this drug is itching/burning. Here I list few reviews:

1. "I honestly wish I was going through childbirth because at least they give you a ton of drugs. THIS PAIN IS UNBEARABLE!! I wouldn’t wish this on my worst enemy... I want to itch so bad that I feel like a dog!"
2. "Omg! I am going crazy... I’m glad to see i’m not the only one... I put the cream in around 10pm and almost immediately it started itching... It’s now 12am and it has not let up at all.. I wish I could scratch my insides it itches so bad.. I got the 3 day but there is no way I am going to put myself through another night of this... I’m thinking a nice hose down of my insides might make me feel better... Wish me luck"
3. "This is the side effect: Imagine a raft of fire ants hunting you down and taking up residence inside your vagina. That’s the best way to describe the feeling I got less than a minute after using this product. The pain lasted 2 hours, until I washed out most of the cream. Only did one day of the 3-day treatment course, then went to see a doctor who prescribed me an oral anti-fungal medication. I’ve actually used this successfully in the past with minimal side effects. But this time...nope. If you are a person who has sensitive skin (or if you’re already feeling irritated from the infection) I would not recommend this. The copay at the doctor’s office for a pill was 1000% worth it.'"

The next low-rated drugs are Belsomra and Suvorexant which are used for insomnia, Dextromethorphan for cough, Movantik and Naloxegol for constipation, Metoclopramide for anti-sickness, Cefdinir is an antibiotic and Depo-Provera for birth control and menopause.

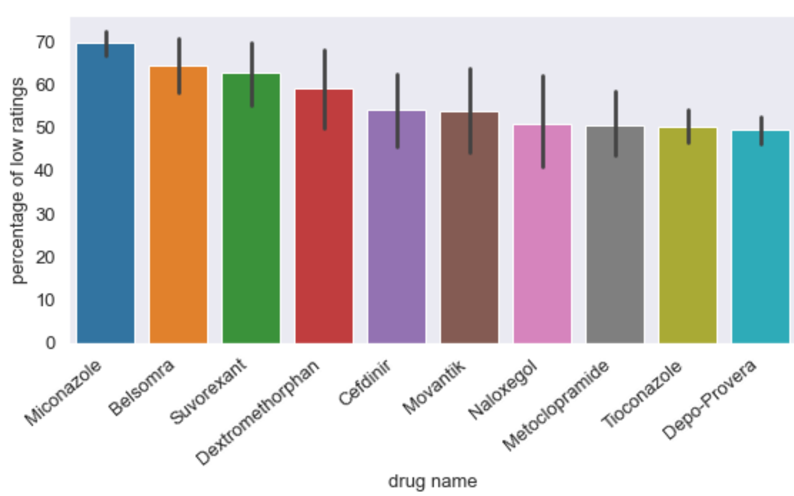

Fig 3.1
Top low-rated drugs

3.2 Least low-rated drugs

The least low-rated drugs are: Percocet; pain-killer, Carisoprodol; muscle relaxant, Adipex-P and Phentermine; weight loss, Magnesium citrate; constipation, Valium and Diazepam are sedatives; anxiety, muscle spasm and seizures, Methadone and Fentanyl are narcotics; treat pain, Suboxone; opiate addiction.

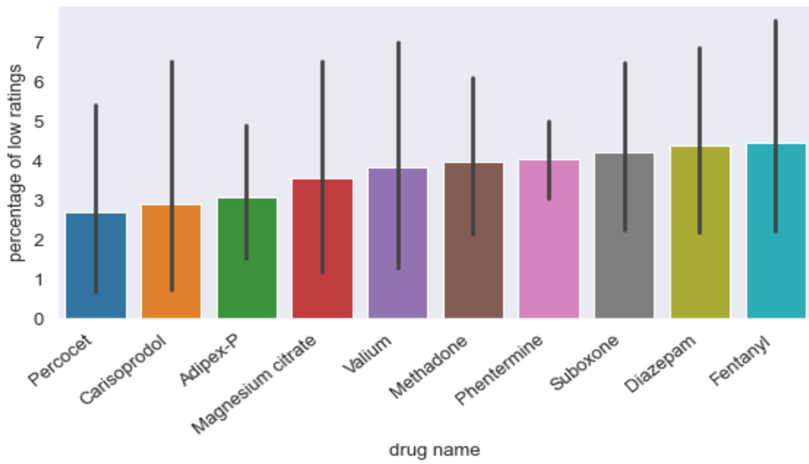

Fig 3.2
Least low-rated drugs

4 Classification/Predict the low-rating drugs based on the review.

4.1 Preprocessing

The pre-processing steps for the review column before text vectorizing is as follows:

1. Lower casing all the text
2. Removing digits
3. Removing punctuation
4. Striping the text
5. Removing drug names from the reviews.

These steps are summarized in a function called preProcess and is passed to the text vectorizer.

4.2 Classification

We use two classification methods in this project for predicting low-ratings based on the reviews: Logistic regression and Multinomial Naive Bayes. Grid searching and 5-fold validation with tfidfVectorizer for Logistic Regression returns the best roc-auc score of 0.93 on both train and test data sets. For the Multinomial Naive Bayes method, Grid searching for minimum document frequency(min_df) and alpha with 5-split cross validation with tfidfVectorizer returns the best min_df =1 and best alpha=0.01. The best roc-auc score on the both test and train sets is 0.95.

Therefore, our best model is Multinomial Naive bayes (MNB) with min_df=1 and alpha=0.01. We choose this model for threshold tuning and further investigation. The classification report table 4.1, for MNB model shows 90% accuracy. The recall and f1 score for low-rated reviews, class “1”, are lower than high-rated reviews, class “0”, because the data is imbalanced.

	precision	recall	f1-score	support
0	0.9	0.98	0.94	41685.0
1	0.89	0.62	0.73	11786.0
accuracy	0.9	0.9	0.9	0.9
macro avg	0.9	0.8	0.84	53471.0
weighted avg	0.9	0.9	0.89	53471.0

Table 4.1
Classification report for Multinomial Naive Bayes method

4.3 Threshold tuning

Threshold tuning on our best model by maximizing f-score returns the best threshold=0.253 with maximum f-score=0.757.

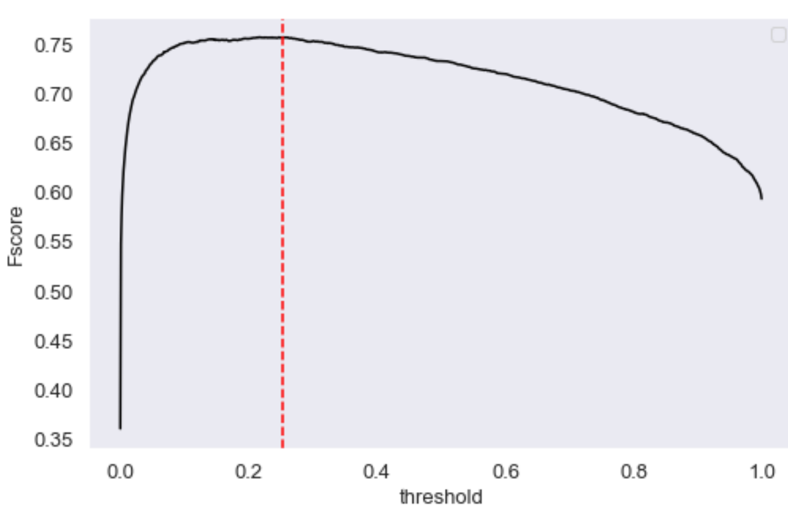

Fig 4.2
F-score vs threshold
The red line shows the best threshold which maximizes the f-score

Using the best threshold of 0.253, returns the new classification report below, Table 4.3, which shows an increase in recall and f1-score for class “1.”

	precision	recall	f1-score	support
0	0.92	0.95	0.94	41685.0
1	0.81	0.71	0.76	11786.0
accuracy	0.9	0.9	0.9	0.9
macro avg	0.87	0.83	0.85	53471.0
weighted avg	0.9	0.9	0.9	53471.0

Table 4.3
Classification report after threshold tuning.

4.4 Most predictive words

For the purpose of finding the best predictive words we choose the minimum document frequency (min_df) of 50 in order to eliminate the words that appear in very few documents. The most predictive words (good words!) and the least predictive words (bad words!) for low-rating reviews are listed in the table 4.3 below.

The good words are as expected since the top eight have negative polarity, and some of the bad words like ‘eases’, ‘proud’, ‘changer’, ‘saver’, ‘believer’, ‘obsessing’ are as expected since they have positive polarity. The birth control drugs have the highest number of reviews in this data set (section 2.4). This might explain the appearance of ‘cum’ and ‘precum’ as least predictive words for low-rated drugs.

	good words	low-rating probability	bad words	low-rating probability
1	poison	0.87	eases	0.01
2	banned	0.85	cum	0.01
3	garbage	0.81	drawback	0.01
4	enemy	0.78	proud	0.01
5	trash	0.78	changer	0.01
6	waste	0.74	saver	0.01
7	rubbish	0.72	believer	0.0
8	ants	0.72	obsessing	0.0
9	blisovi	0.71	precum	0.0
10	pulmonary	0.7	presentations	0.0

Table 4.3
The most predictive words (good words) and the least predictive words (bad words) for low rating drugs.

This same analysis can be applied to individual drugs as well. The top low-rated drug is ‘Miconazole’ used to treat vaginal yeast infection with 70% low rated reviews, and the least low-rated drug is ‘Percocet’ used as a painkiller with 27% low rated reviews. The common side effect and complaint about ‘Miconazole’ is itching and burning. Our model found the most predictive words for low-rating reviews of ‘Miconazole’ to be: literally, buy, horrible, wish, stuff, awful, worst, life, vagina and trying.

4.5 Mispredicted ratings

A few samples of mis-predicted high rating reviews shows that the review started negative by discussing a treatment that was not working and then discussing the drug that the patient is currently using. This caused the review to be mispredicted as low-rating. Two samples of mispredicted high rating:

1. "Took Cipro and Bactrim during Chemo and Radiation. Did not work, may have been because of the Chemo and Radiation, just not sure. I am now on my 3rd of 5 day Levaquin dose and passed my AZO UTI over the counter test this morning, yea. Two months of a UTI is a bit much, although, I now know what works for future."
2. "I developed cystic acne, I’m guessing as a result of hormones, stress, and genetics. My doctor and I changed what we were using during the second visit, since we were doing some trial and error (figuring out what works). Since my skin is not sensitive, we went from a combination of doxycycline, tretinoin, and clyndimicine (idk how to spell it) to a combination of tretinoin of higher strength, erythromycin/BP gel, and minocycline. My skin has cleared up in a matter of 2 months with the occasional pimples and some acne scarring."

And few samples of mis-predicted low rating reviews, “Not Good” and "Not good enough," show that the model did not use the “not good” 2-gram as a predictive negative word.

5 Takeaway and future research

In this project, we built multiple classification models that predict the low-rated drugs based on the reviews. The best model, Multinomial Naive Bayes, achieved a roc-auc score of 95%, on both the train and test data sets, and an accuracy of 0.9, which are considerably high. This result was achieved by defining a function which preprocessed the text before passing it to the text vectorizer, by grid searching the model and by 5-fold cross validating the data.

This could be used by pharmaceutical companies to find out what things people do and don't like about a new drug that's been released, learn about the side effects as well as get an idea for the overall response to a new drug.

In this project, we built a model to predict the low rated drugs. However, identifying high rated drugs is crucial for pharmaceutical companies too. By identifying high rated drugs, pharmaceutical companies can adopt a business strategy that increases their profit, by investing and allocating their resources on more popular drugs. For future research, it would be interesting to study the transferability of the model to new datasets accessible from different websites.