How Drug Reviews Can Aid in Prescription of Appropriate Medications: Sentiment analysis and feature selection as methods of extracting patterns

Ivette Planell-Mendez[1]

[1] Princeton University, Princeton Neuroscience Institute, Princeton, NJ

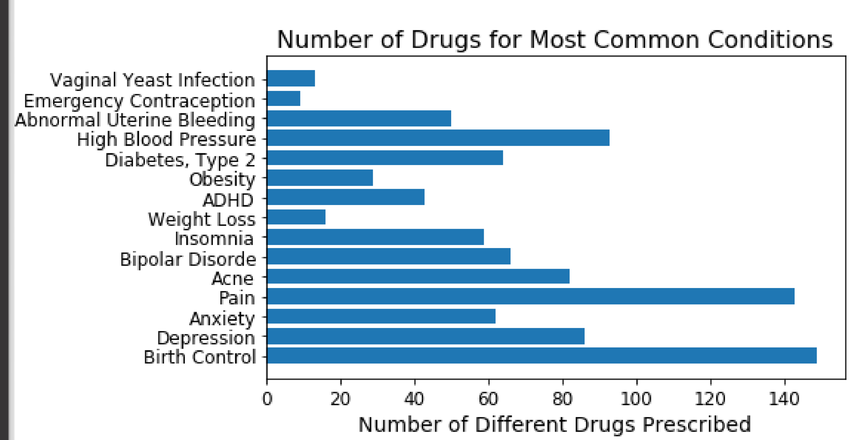
ABSTRACT

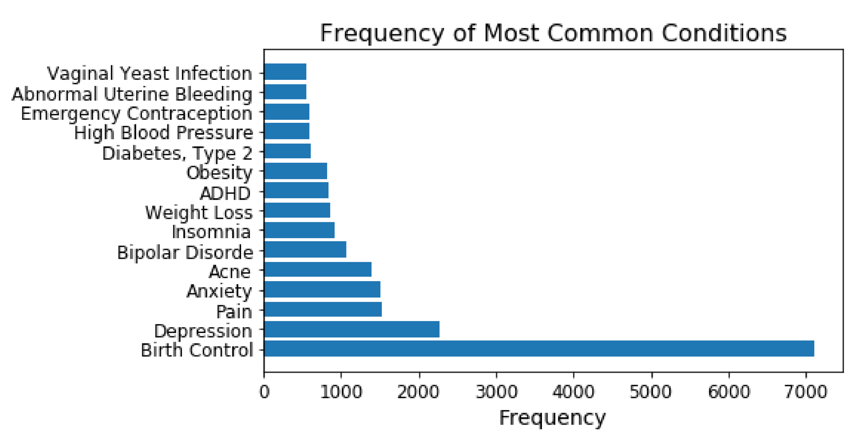
This study we aim to develop analysis techniques that can be applied to publicly available drug review data in order to improve methods for matching patients to drugs that will treat their symptoms while minimizing side effects. We use sentiment analysis to extract patterns in the reviews, classification to predict the patients condition based on the drug, and regression to predict the rating of the drug based on the reviews.

DRUG REVIEW DATASET

The Problem: The pharmaceutical market, especially in the case of certain conditions, is oversaturated with drug options. Patients must go through a lengthy period of trying different drugs until they find one which minimizes side effects.

The data set considered consists of 40,000 publicly available drug reviews. Each entry in the data set consists of the name of the drug being reviewed, the condition for which that drug is taken, a written review of the drug, a ratings on a scale of 1-10, and the number of users that considered the given review useful. The data set sheds light on the problem that patient's face when choosing a the correct drug. The data set contains 2386 different drugs for 679 conditions. That is a ratio of 3.5 drugs per condition; however, certain conditions have many more drug options and these conditions are some of the most common in the data set. Birth control, which is by far the most common use in the data set, has over 140 drugs to choose from.





Hypothesis: Machine learning methods can be used to extract patterns and information from publicly available drug reviews to better match patients to specific drugs and to lessen the burden on patients.

When the reviews are separated by the 1-10 ratings provided by users, we can look at which words are most common for the distinct cases. We see some differences; however, there is an overwhelming amount of repeated words across the different categories. This promotes a need to employ various machine learning methods to extract patterns and select the most predictive features.

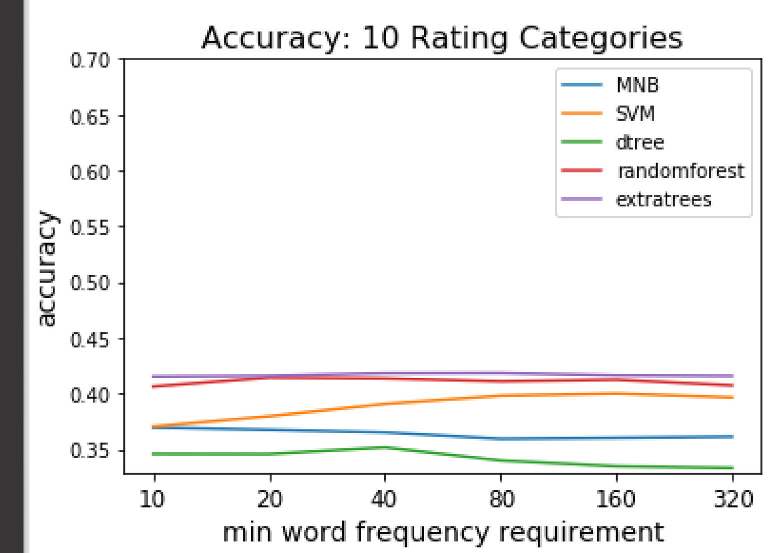
SPECIFIC QUESTIONS ADDRESSED

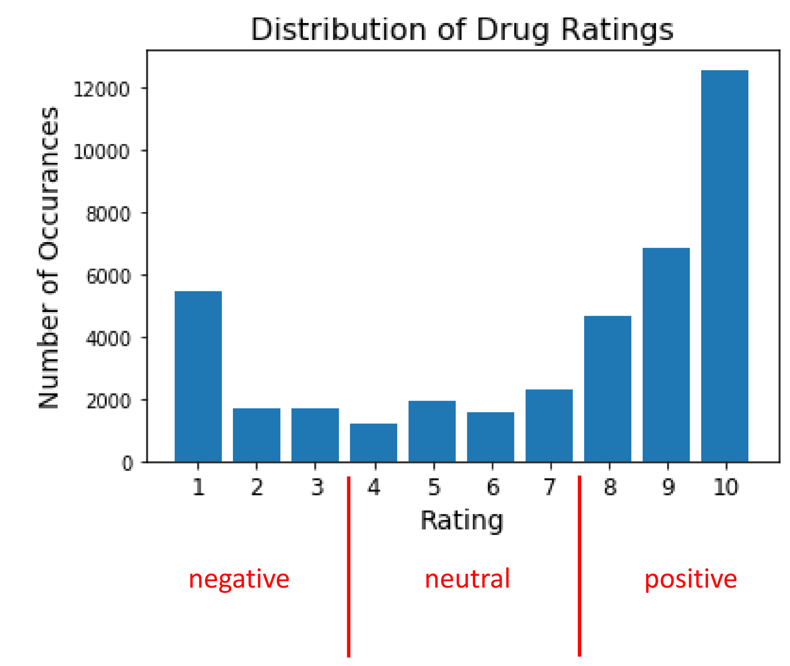
1. How does the minimum repetition count for use in bag of words representation affect performance in predicting ratings?
2. How does reducing from ten rating categories to three (positive: 8-10, neutral: 4-7, negative: 1-3) affect performance?
3. Which methods of feature selection work best with this data set?
4. Can we use regression to predict the user's condition based off of their review?
5. What topics can be extracted from the reviews using unsupervised dimensionality reduction techniques?

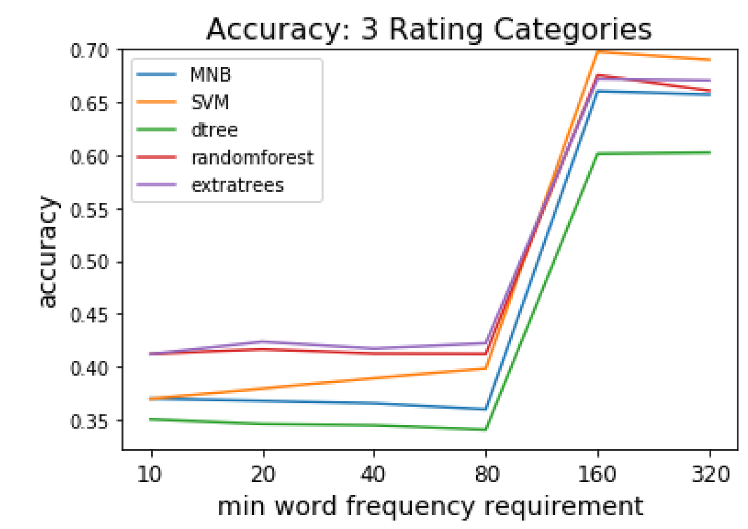
PREDICTION OF DRUG RATINGS

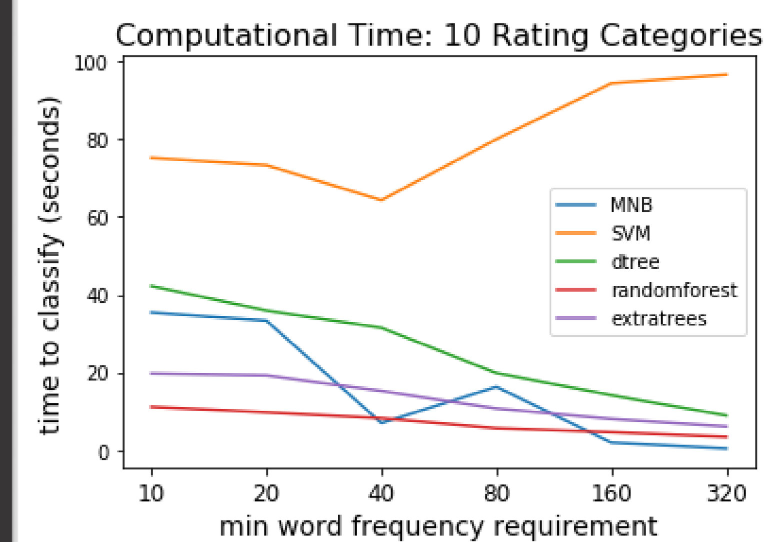
In addition to leaving a written review users also rated the drugs on a 1 – 10 scale. Interestingly the distribution of numerical ratings given were predominantly high or low, with the fewest in between. This presents a challenge when extracting patterns from the written reviews as there are not an equal number of reviews per rating category; therefore, some categories are likely to be harder to predict than others because there is less information about them. To try to deal with this issue we grouped review categories into three groups: negative (1-3), neutral (4-7), and positive (8-10).

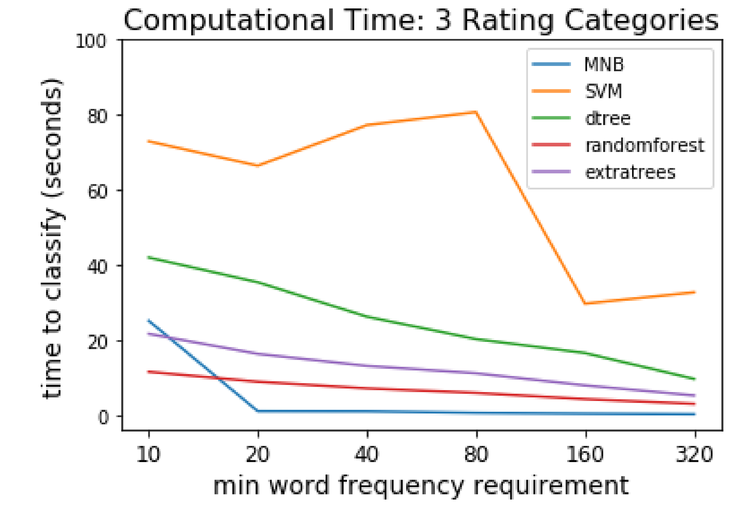
We also looked at how imposing different minimum word frequency constraints when forming the bag of words representation affected performance. We saw that this didn't increase performance in the 10 category rating case, but made a big difference in the case of the three rating class condition.



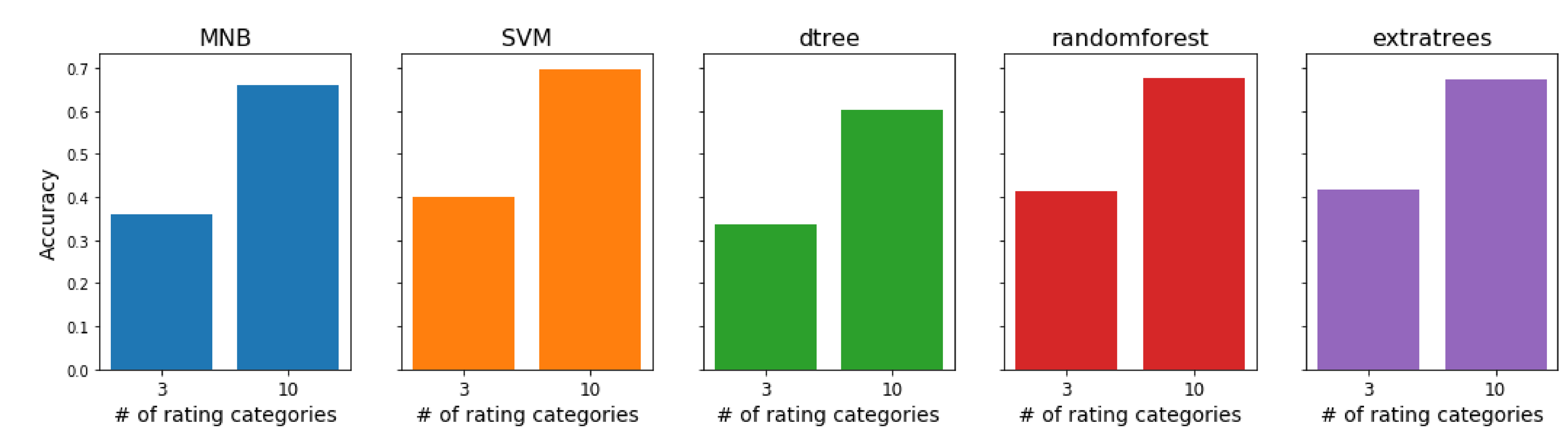




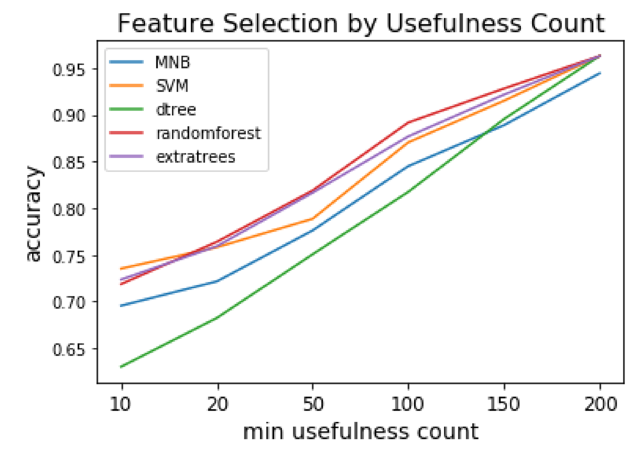




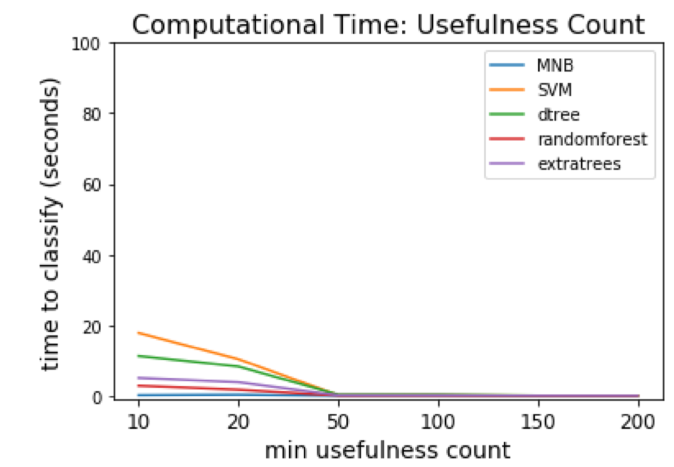
Since the bag of words representation which required a minimum word frequency of 160 performed best, we directly compared the performance across the different classifiers for both the three class (negative, neutral, positive) condition and for the 10 rating class (1-10) condition. There was much better performance across the board for the three class case. Performance almost doubled in most cases.



Users of the websites where the drug reviews were collected from could flag a review as useful whenever they considered it useful (we call this the usefulness count). We used the usefulness count in order to do feature selection and see if it would improve performance. Usefulness count is intuitively a good feature selection metric because the users are telling you which reviews were most informative.



Accuracy for Min. Usefulness Count = 200	
MNB	0.944
SVM	0.963
dtree	0.963
randomforest	0.963
extratrees	0.926



ACKNOWLEDGEMENTS

Special thanks to the assistants in instruction for their guidance during precept, to Dr. Engelhardt for informative lectures, and to the Princeton Neuroscience Institute, whose funding and support made the completion of this work possible.

REFERENCES

[1] UCI. Uci machine learning repository, 2019.
[2] F Graber, S Kallumadi, H Malerg, and S Zaunsedr. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning.
Proc. of the 2018 International Confer-
ence on Digital Health

ivettep@princeton.edu