Medical Drug Review Analysis

Simon Reif
Matrikelnummer: 5617379
simon.reif@student.uni-tuebingen.de

Dejana Mandic
Matrikelnummer: 6024572
dejana.mandic@student.uni-tuebingen.de

Abstract

User reviews of medical drugs are increasingly available and provide additional information to those mindful of selection biases. For this article we explored a public dataset of drug reviews to present the following: a ranking of medical conditions and drugs according to the usefulness of their reviews, a way of assessing the importance of word occurrences in a review for predicting the associated rating, and a comparison of popular sentiment analysis models. Code is available at https://github.com/dmandic17/medical-drug-review-analysis.

1 Introduction

The internet era is a period in the information age in which communication and commerce via the internet became a central focus for businesses, consumers, government, and the media. Since we have long been in the internet era, we are aware of the fact that people inform themselves about all types of products through other people's comments and reviews. Both for readers of reviews and data analysts it is important to keep in mind that reviews don't come from a representative sample of drug users, but are self selected; they don't replace clinical trials. However, as we are all aware, almost all medical drugs can cause side effects. These side effects are supposedly rare but sometimes can lead to serious consequences, so the key is to stay informed. This makes the reviews a database from which people can learn through others' experiences.

In this project, we are examining the UCI ML Drug Review Dataset (Li, 2019), which contains a list of medical drugs along with user reviews and ratings. We are exploring for which medical conditions and medications is the highest number of people informed through other people's reviews and providing an explanation why that might be the case. Furthermore, we explore the importance of certain word counts for predicting the rating of a review using linear regression models. In this case, we have limited the reviews to the one medical condition: *depression*, as it is the condition with the highest number of people that find the reviews for its medications useful. Moreover, we are comparing a few models for sentiment analysis of the reviews. Sentiment Analysis is the identification of the sentiment associated with a sentence, phrase or an entire document. Since we are dealing with reviews of drugs in this dataset, we would certainly be interested in understanding the sentiment associated with these reviews and try to see which drugs have positive reviews and which ones have negative reviews.

2 Experiments and Results

In this section, we describe how we filtered the dataset and what are the medications and medical conditions with the highest number of useful reviews. Further, we try to predict the ratings from word occurrences using linear regression and explore the correlation of these occurrences with the ratings. Finally, we compare three different models on the task of sentiment analysis: Random Forest (Breiman, 2001), Multinomial Naive Bayes (Kibriya et al., 2005) and XGBoost (Chen and Guestrin, 2016).

Project Report for *Data Literacy* 2021/22

2.1 Data Description

The data provided by Li, 2019 contains *reviews* and *ratings* of medical *drugs* for a specific applications (*condition*) as well as their *date* and how many user users marked them useful (*useful count*). It is split into a training set of approximately 160,000 and a test set of 53,000 reviews.
Since many conditions appeared only once or were nonsensical, we kept only entries, whose condition occurs more than 500 times in the data set; reducing the training set to 130,000 and to test set to 43,000 entries.

2.2 Useful reviews of medical drugs

We wanted to examine whether the top ten medical conditions ranked by how many people found the reviews for them useful (*usefulCount*) overlap with the top ten medications with the most useful reviews. The appropriate plots can be seen in Figure 1. As expected, the top conditions (by *usefulCount*) are common, every-day conditions, and top medications (by *usefulCount*) are the medications for some of these conditions. So, from these results, we can see that the overlap exists and that the most useful reviews are for conditions that are either very common and have a lot of different medications (e.g., pain, high blood pressure etc.) or have medications with frequently occurring side effects (e.g., birth control, depression, anxiety etc.).

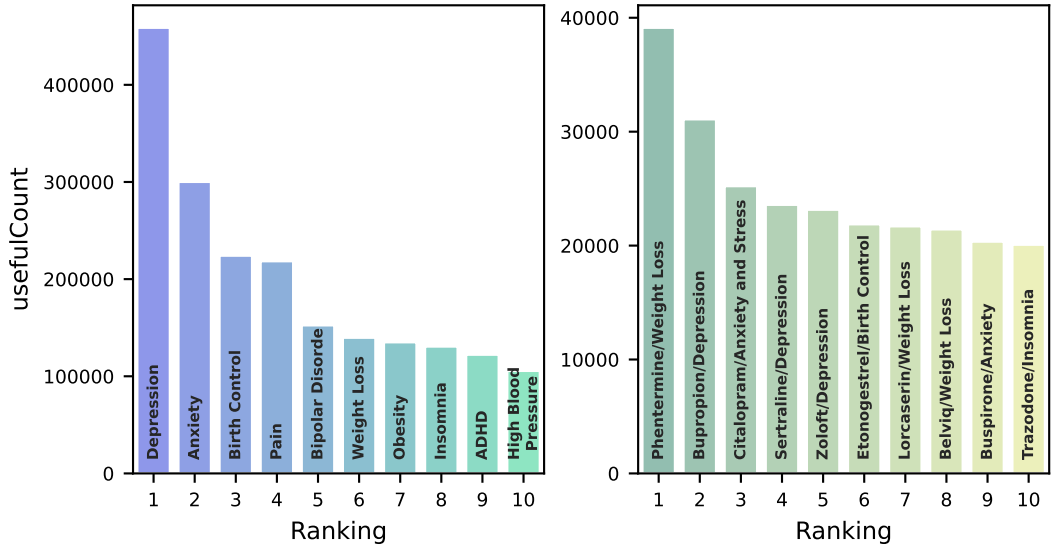

Figure 1: Top 10 medical conditions by how useful the people found the reviews for their drugs *(left)* and top 10 drugs along with medical conditions they are used for according to the same criteria *(right)*.

2.3 Predicting ratings from word counts

The simplest way of processing text is to count word occurrences and collect them in a vector, a "bag of words". In the following, we explore the importance of certain word counts for predicting the rating of a review. Although ratings are discrete, their semantics makes regression a more natural choice than classification. Predicting a rating of 8 when the true rating would be 2 is worse than predicting 3.
Figure 2 shows the performance of linear regression models trained with word vectors containing one through fifty of the most common words in reviews of *depression* medication. Performance is measured with the coefficient of determination R^2 (Wright, 1921), which has the advantage of accounting for the unconditional distribution of ratings. The best constant classifier has an R^2 value of 0. As expected, performance goes up with more information - we observed the same with more flexible models like decision trees, who reach an R^2 test score of approximately 0.7 with 50 input words.

The slope of the graph gives an indication of the importance of a word for the prediction, in particular, it shows the importance given preceding words. A steep slope here means that a word is important overall, but a shallow slope does not necessarily mean that it is not important. In this particular case, different permutations showed similar slopes at the same words.
In the background of Figure 2, the bar plot shows a measure of the absolute correlation between word occurrences and rating. Since review lengths are different for different ratings, we normalized word occurrences with text length. All important additional words for the regression have high correlations with ratings, but a high correlation doesn't mean a word is an important addition, because of correlation between words.

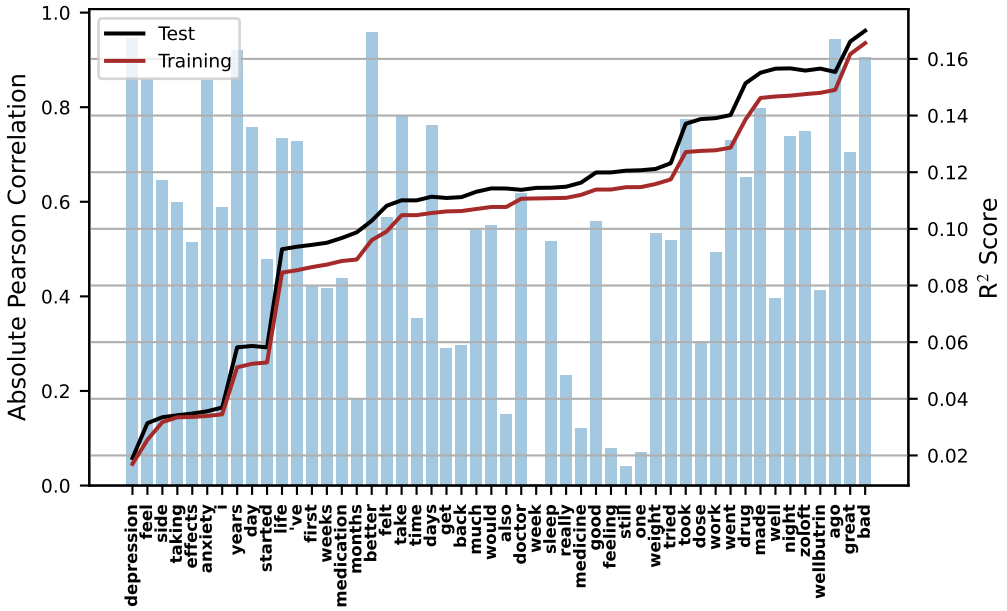

Figure 2: Lines show R^2 score of different linear models trained with word vectors containing words up to the ones labeling the x-axis; Bars show the absolute value of the Pearson correlation coefficient of word counts and ratings. Data is limited to the reviews for depression medications.
The test score is a little better, which is likely a coincidence as the difference is very small and decision trees didn't show this behavior.

2.4 Sentiment analysis of reviews

Sentiment analysis is an important task in all types of reviews as it helps in better understanding the emotion in the review and provides a possible direction for product improvements. This data is especially good for comparing the performance of different models on the task of sentiment analysis as it contains both the *reviews* for medical drugs and the *rating* from the user who wrote the review. We have separated the ratings in three different sentiment classes where *class 0 = negative* is for $rating < 5$, *class 1 = neutral* for $rating = 5$ and *class 2 = positive* for $rating > 5$. For the purpose of this task, we first evaluated all reviews using TextBlob Sentiment Analysis tool (?), using both filtered (removing stopwords and punctuation) and unfiltered reviews, and found the correlation between sentiment and grade is lower when using filtered data. The possible explanation for this is that word *"not"* is among the stopwords, and when it is eliminated the sentiment might be predicted wrongly.
We have also compared the performance of three standard models: Random Forest (Breiman, 2001), Multinomial Naive Bayes (Kibriya et al., 2005) and XGBoost (Chen and Guestrin, 2016). All models were trained on test data and evaluated using test data where target classes were calculated as described previously. Since a classifier requires numeric input, the reviews were first converted to vectors using TF–IDF (term frequency-inverse document frequency) which is a term weighting scheme commonly used to represent textual documents as vectors (Sammut and Webb, 2010).

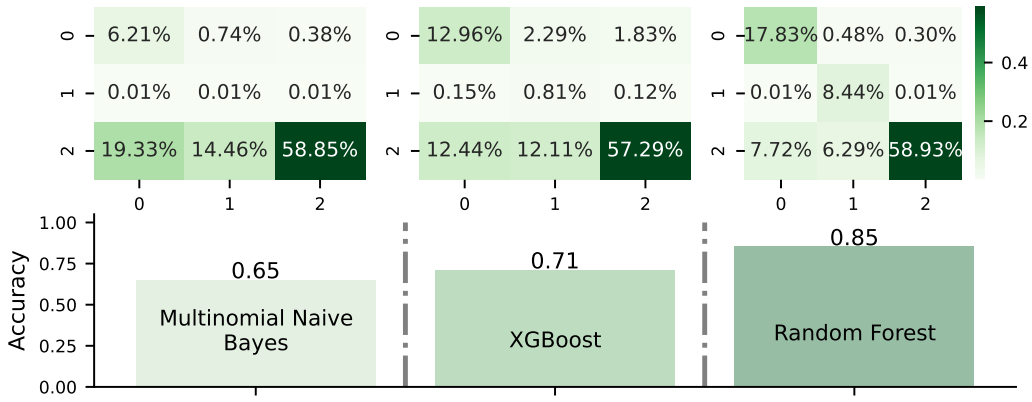

Figure 3: Accuracy plot and Confusion Matrices for MNB, XGBoost and RF models where $0 = negative, 1 = neutral$ and $2 = positive$ sentiment.

The accuracy plot along with confusion matrices for all models can be seen in Figure 3. We can see that Random Forest outperforms all other models (with 85% accuracy), even without any parameter tuning. We can tune the parameters of our classifier and improve the models' accuracy. Another important thing we can conclude is that all models seem to be better at predicting *positive (class 2)* sentiment than *negative (class 0)* and *neutral (class 0)* sentiment. This is a consequence of the unbalanced number of samples of different classes - the largest number of samples in the training set comes from *positive (class 2)* sentiment. In the future, we could try to balance the dataset using data augmentation: for example, changing some words with their synonyms because that would retain the sentiment of a review. This data augmentation could result in improving the accuracy.

3 Conclusion

We have shown a way of assessing the importance of word occurrences in a review for predicting the associated rating. Furthermore, we have also seen that the accuracy of commonly used sentiment analysis models is on an acceptable level for this task but could further be improved by balancing the dataset because there are more positive reviews than negative ones. Identifying the sentiment of a review, and knowing which words might be important for the rating can help us make an attempt at predicting the rating that the user would give to the given product and also help the pharmaceutical companies in better understanding the reaction of their customers.

References

Breiman, L. (2001). In: *Machine Learning* 45. DOI: 10.1023/A:1010933404324.

Chen, T. and C. Guestrin (2016). "XGBoost". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. DOI: 10.1145/2939672.2939785.

Kibriya, A. M., E. Frank, B. Pfahringer, and G. Holmes (2005). "Multinomial Naive Bayes for Text Categorization Revisited". In: *AI 2004: Advances in Artificial Intelligence*. Ed. by G. I. Webb and X. Yu. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 488–499.

Li, J. (2019). *UCI ML Drug Review dataset - Version 2*. Available at https://www.kaggle.com/jessicali9530/kuc-hackathon-winter-2018, version 1.6.0.

Sammut, C. and G. I. Webb (2010). "TF–IDF". In: *Encyclopedia of Machine Learning*. Boston, MA: Springer US, pp. 986–987. DOI: 10.1007/978-0-387-30164-8_832.

Wright, S. (1921). "Correlation and causation". In.