Capstone Project 2: Final Report

Mehmet Erdoğdu

Nowadays the narrative of a brand can be largely affected by opinions of customers (and strangers) posted on the web. For this reason, companies are constantly looking out across Blogs, Forums, and other social media platforms, etc. to check sentiment for their various products (as well as competitor's products) to learn how their brand resonates in the market. Furthermore, many symptoms and side effects of drugs are hidden within online reviews, which can be harnessed to improve both the drugs themselves and the effectiveness of their prescription. This is all essential information for pharmaceutical companies and medical institutions.

In this project, I built a model to predict the ratings of the drugs based on the reviews, and analyzed what features of different drugs are most often associated with positive and negative sentiment.

The Dataset

I have used UCI Machine Learning - Drug Review Dataset. I acquired the data set through Kaggle website, which can be found at this link:
https://www.kaggle.com/jessicali9530/kuc-hackathon-winter-2018

The dataset has 7 columns with a total 161297 entries.The column names as follows:

uniqueID:Unique ID
drugName:Name of drug
condition:Name of condition
review:Patient review
rating:10 star patient rating
date:Date of review entry
usefulCount:Number of users who found review useful

Exploratory Data Analysis

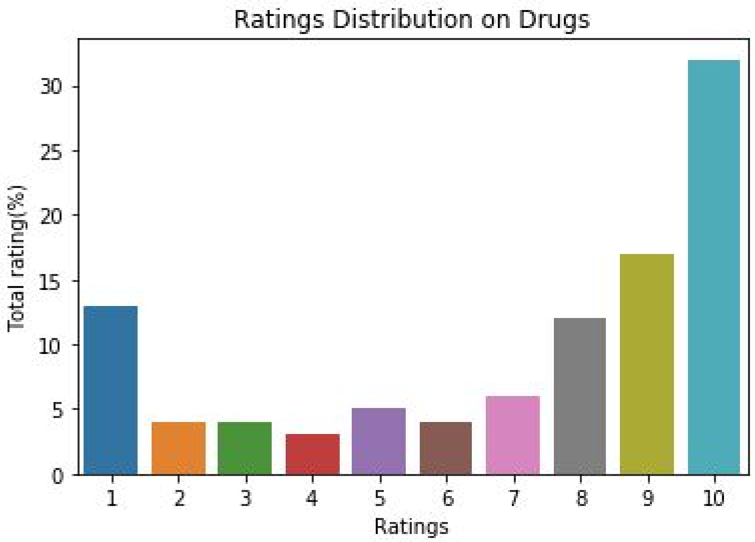


Figure 1 Rating vs number of reviews

Each review entry has the text of the review along with the user rating on a scale of 1-10. Above, you can see the distribution of the different ratings. 10 is the most common rating by reviewers with more than 50% of the all reviews, with 4 being the least common.

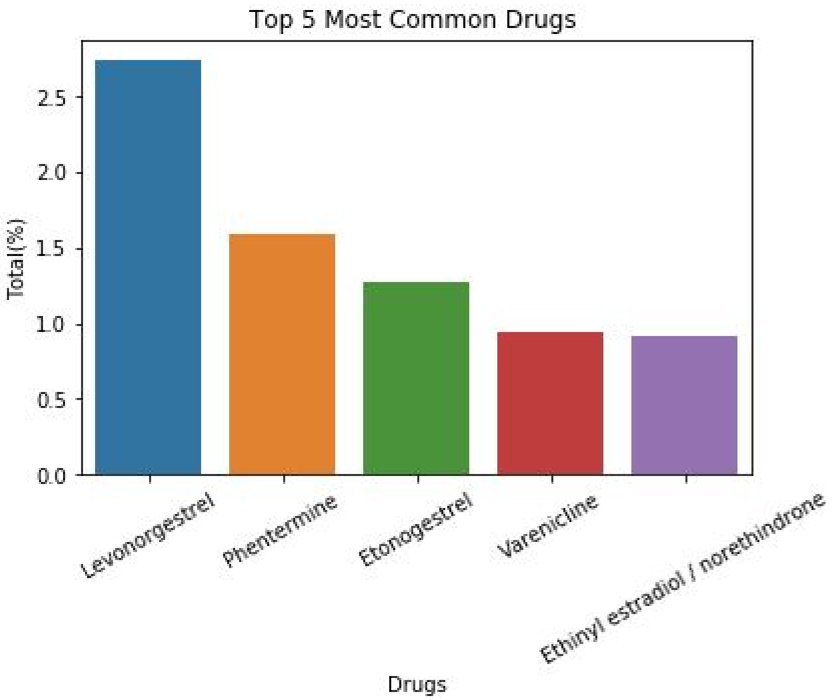


Figure 2 Top 5 most common drug name

I have 3436 unique drug names in the data set. Levonorgestrel is the most common drug reviewed in the dataset. Above you can see the top 5 most common drug names.

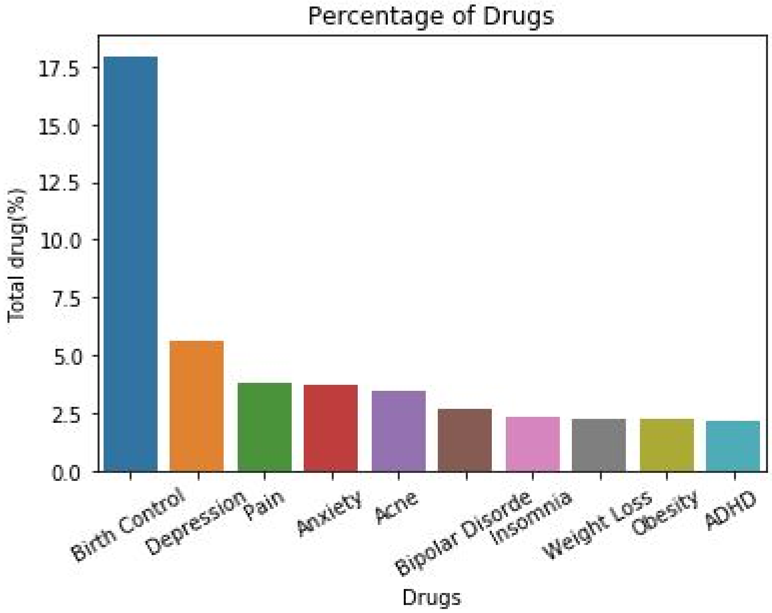


Figure 3 Top 10 most common conditions

Total number of unique conditions in this data set is 885. Birth control, Depression, pain, anxiety and acne are the top 5 conditions. Above you can see the top 10 conditions.

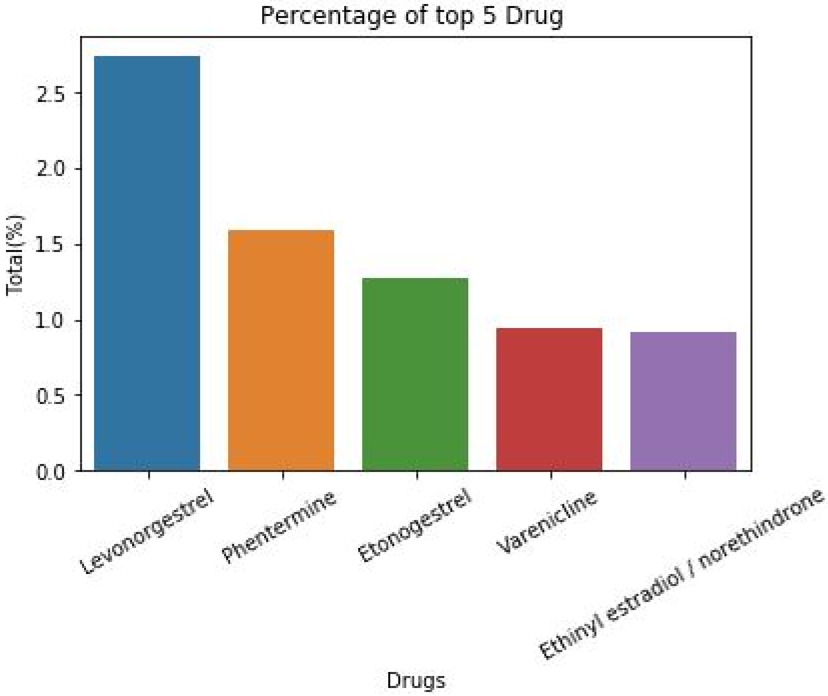


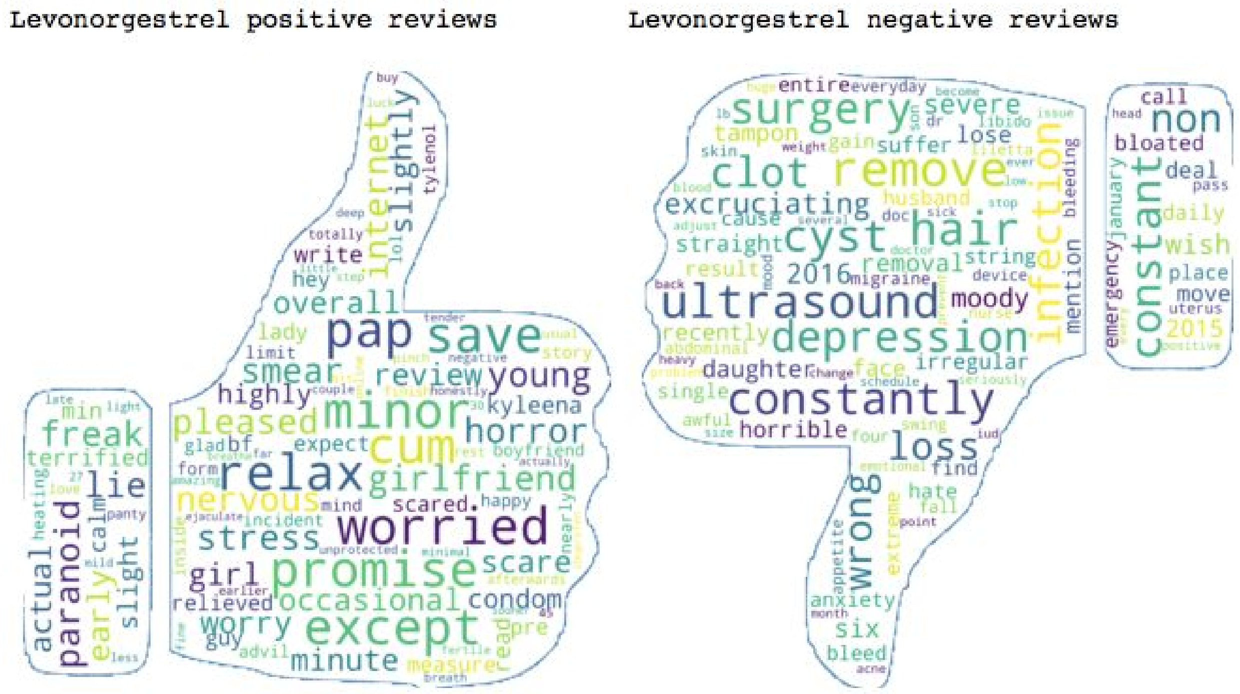
Figure 4 Top 5 Drugs rated 10

Above you can see the percentage of top 5 drugs reviewed in the data set. Levonorgestrel, Etonogestler and Ethinyl estradiol/norethindrone are birth control medications, Phentermine is a weight loss medication and Varenicline is a medication for smoking cessation.

I calculated the t-test for the means of two independent drug names within top 10 rated drugs. The t-test shows that relationships between ‘Sertraline-Levonorgestrel’ , ’ Nexplanon-Etonogestrel’ , ’ Ethinyl estradiol/norgestimate – Etonogestrel’ , ’ Ethinyl estradiol/levonorgestrel – Etonogestrel’ , ’Ethinyl estradiol/norgestimate – Nexplanon’ , ’Ethinyl estradiol/levonorgestrel – Nexplanon’ , ’Ethinyl estradiol/levonorgestrel - Ethinyl estradiol/norgestimate’ are not statistically significant since p-value is greater than 0.05

I created a function called ‘mycloud’. It takes 2 parameters; data, and ‘positive’ or ‘negative’ words. Within the function, I created an instance of the *CountVectorizer* class. And with the fit_transform function, Mycloud learned vocabulary from documents and encoded each document as a vector. An encoded vector is returned with a length of the entire vocabulary and an integer count for the number of times each word appears in the document.

After creating a term-document matrix with CountVectorizer I trained a predictive model on the matrix. Since all features are integers I used Naive Bayes classifier to train and test my model. I created an identity matrix the length of the vocabulary and used my model to predict each word on the vocabulary and got a list of probabilities for each word with predict_proba. I then sorted for most/least probable words for any given class.



Levonorgestrel positive reviews

Levonorgestrel negative reviews

Figure 5 Levonorgestrel positive and negative reviews

The most common drug in the dataset is Levonorgestrel, which is a form of birth control. Above you can see the 100 most predictive words for either positive or negative sentiment for this drug. 100 positive and negative words.For positive sentiment 'worried', 'minor', 'except', 'pap', 'cum', 'relax', 'promise', 'save', 'girlfriend', 'lie' are most probable. For negative sentiment 'remove', 'ultrasound', 'hair', 'constantly', 'cyst', 'surgery', 'depression', 'constant', 'infection', 'clot' are most probable.

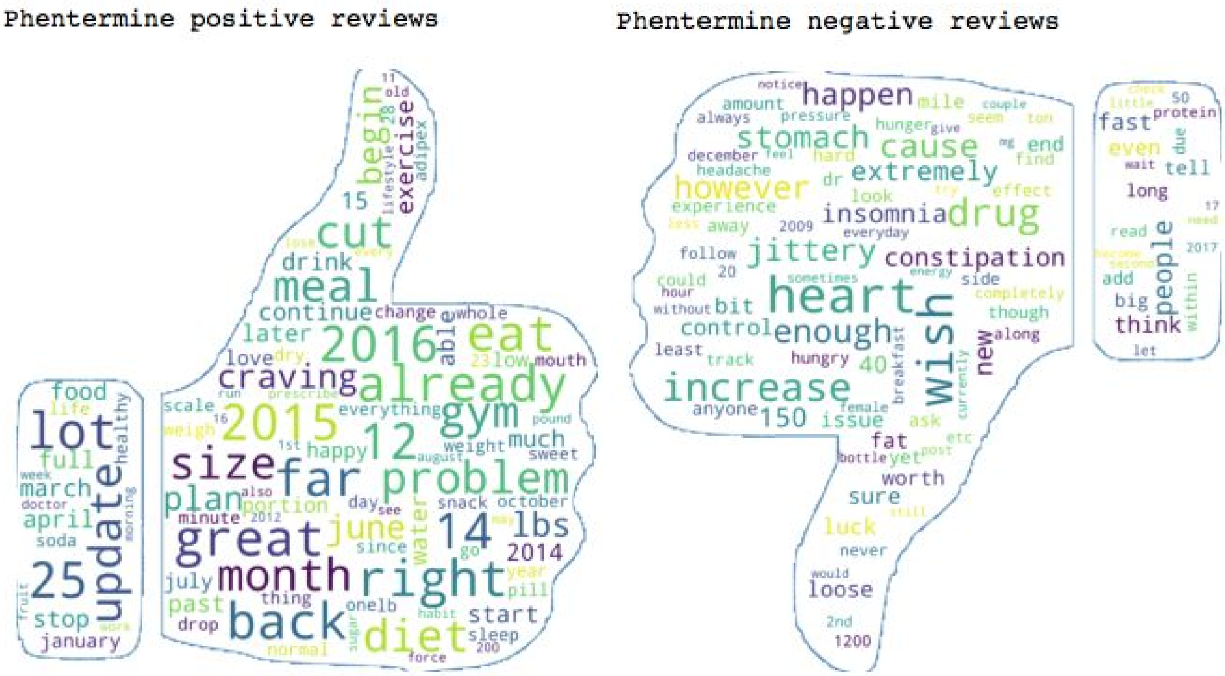


Figure 6 Phentermine positive and negative reviews

The second most common drug in the data set is Phentermine, which is used to exercise to treat obesity.Above you can see the 100 most predictive words for either positive or negative sentiment for this drug. 100 positive and negative words.For positive sentiment 'great', 'far', 'eat', 'lot', 'right' are the most probable. For negative sentiment 'heart', 'wish', 'increase', 'drug', 'cause', 'enough', 'however', 'stomach', 'jittery', 'happen' are most probable.

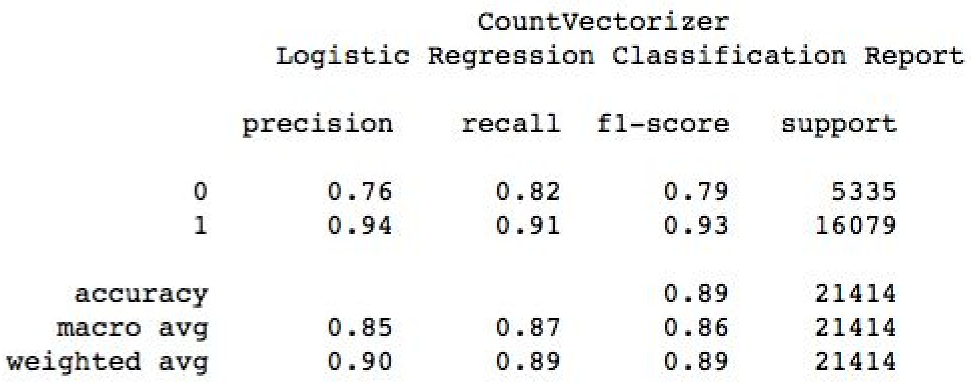
Machine Learning

I began by doing an 80/20 split for my train and test sets - the test set remained untouched until the final model was selected and tuned. I then used a Pipeline object to chain estimators so I could gridsearch not only the parameters of the models but also for CountVectorizer and TFIDFVectorizer to optimize for minimum document frequency and maximum document frequency. I used balanced accuracy as my metric to determine the best parameters for each model, grid searching over 5-fold cross-validation. The resulting best parameters for each model are below:

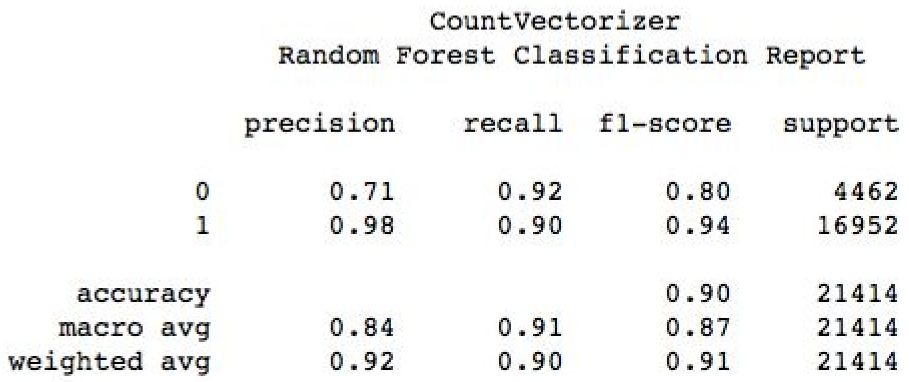
Model	TFIDF Parameters	TFIDF Roc-Auc Score	CountVectorizer Parameters	CV Roc-Auc Score
Logistic Regression	C : 50 Min_df : 1 Max_df : 0.95	0.856	C : 6 Min_df : 1 Max_df : 0.95	0.857
Naive Bayes	Max_df : 0.8 Min_df : 10	0.682	Max_df : 0.8 Min_df : 1	0.806
Random Forest	Max_depth : 100 Max_features : 0.1 N_estimators : 100 Max_df : 0.95 Min_df : 20	0.823	Max_depth : 100 Max_features : 0.1 N_estimators : 100 Max_df : 0.95 Min_df : 10	0.840

Countvectorizer

I first used Countvectorizer to both tokenize the collection of reviews and build a vocabulary of known words, and then encode new documents using that vocabulary. I performed gridsearch for maximum documentation frequency and received the same score as 0.95 so I assumed maximum documentation frequency as 0.95 for all models. Below you can see the classification report for Logistic Regression and Random Forest Classification.



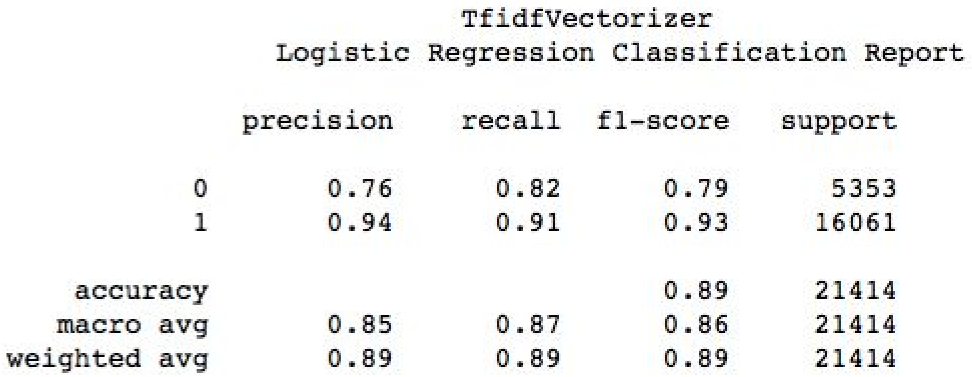
```
                    CountVectorizer
       Logistic Regression Classification Report

              precision    recall  f1-score   support

           0       0.76      0.82      0.79      5335
           1       0.94      0.91      0.93     16079

    accuracy                           0.89     21414
   macro avg       0.85      0.87      0.86     21414
weighted avg       0.90      0.89      0.89     21414
```



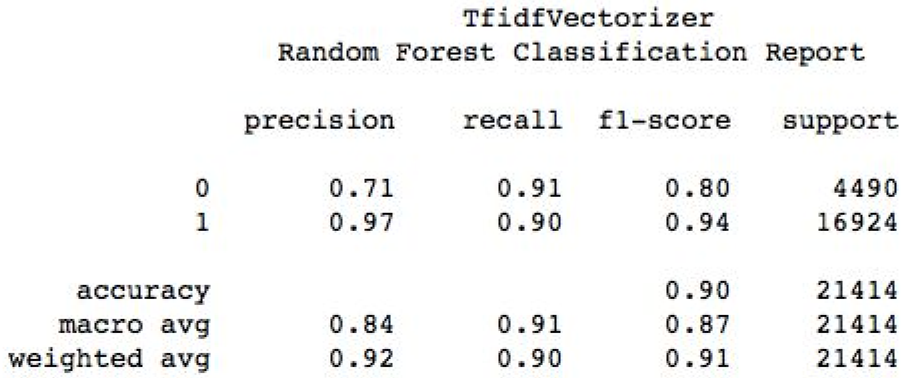
```
                 CountVectorizer
       Random Forest Classification Report

              precision    recall  f1-score   support

           0       0.71      0.92      0.80      4462
           1       0.98      0.90      0.94     16952

    accuracy                           0.90     21414
   macro avg       0.84      0.91      0.87     21414
weighted avg       0.92      0.90      0.91     21414
```

TFIDF

Secondly, I used TFIDF which is Term Frequency – Inverse Document” Frequency to tokenize documents, learn the vocabulary and inverse document frequency weightings, and allow me to encode new documents. Below can be seen the classification report for Logistic Regression and Random Forest Classification.



```
                TfidfVectorizer
     Logistic Regression Classification Report

              precision    recall  f1-score   support

           0       0.76      0.82      0.79      5353
           1       0.94      0.91      0.93     16061

    accuracy                           0.89     21414
   macro avg       0.85      0.87      0.86     21414
weighted avg       0.89      0.89      0.89     21414
```



```
                TfidfVectorizer
     Random Forest Classification Report

              precision    recall  f1-score   support

           0       0.71      0.91      0.80      4490
           1       0.97      0.90      0.94     16924

    accuracy                           0.90     21414
   macro avg       0.84      0.91      0.87     21414
weighted avg       0.92      0.90      0.91     21414
```

Results

After building and testing over 3 models with 2 separate vectorizer, I concluded that Logistic regression has the highest accuracy score with 0.857 with 6% False Positive rate for Countvectorizer and 0.856 with 6% False Positive rate for TFIDFvectorizer.

Conclusion and Further Thoughts

While drug companies get feedback from clinical trials in the development of their products, these are limited in size and scope and online reviews can offer an important next level of feedback on a wider scale. Drug companies may consider modifying drugs based on customer reviews for the next generation production phase. They can keep the positive features based on positive reviews. They can also focus on customer complaints and can make changes to drugs based on negative reviews. This will help the companies produce better products and solve the health issues of customers more efficiently.

Furthermore, aside from the actual effects of the drugs, online reviews can largely affect the narrative and perception of a pharmaceutical product. Understanding how a perceived is essential to properly marketing it. Using the techniques and models developed in this project, pharmaceutical companies can better understand and market their products, and deal with any issues in perception promptly to prevent them from becoming larger issues.

In particular,I built a predictive model that was also able to classify positive and negative reviews with a 0.86 balanced accuracy score. This model could be used directly by these companies. Next steps to improve the model even further would include performing other machine learning models or specific models for specific drugs.