Sentiment Analysis and Condition Prediction on Drug Reviews

Akshay Bhala, Sharvil Vilas Turbadkar, Yeswanth Reddy Velapalem
School of Information Studies, Syracuse University, Syracuse, New York, USA
{abhala, svturbad, yvelapal} @syr.edu

ABSTRACT
Sentiment Analysis is one of the important parts of text mining approaches designed to identify the sentiment polarity within text. The sentiment analysis of healthcare and users' drug experience in particular could provide some insights on how to improve public health and make the right choices. We are also trying to predict the condition of the person based on the reviews by building a supervised classification algorithm. We implement different vectorization techniques and trained the algorithms on top 10 conditions to reduce the complexity of the model. The multi-task learning is further utilized to transfer the helpful domain knowledge from the short text-level drug review collection.

Keywords:* *Text Mining, Sentiment Analysis, Multi-class Classification

1. INTRODUCTION

The drugs in the pharma industry are introduced into the market after conducting multiple experiments and trials under specific protocols and standardized conditions. Still, they might act in different ways in different types of patients. Therefore, it is very crucial for the industries to collect information from the users to improve the effectiveness of the drugs. Online review sites and discussion forums contain a lot of information regarding user experiences over various products which can be used to obtain valuable insights. In this project we examine drug reviews given by various users which contain information related to aspects such as effectiveness of drugs, which make analysis very interesting. Analyzing sentiments concerning the various aspects of drug reviews can provide valuable insights, help with decision making and improve monitoring patient health[1]. In this project we intend to perform multiple tasks such as supervised classification and sentiment analysis over drug reviews with data obtained by crawling online review sites. We plan to perform sentiment analysis to predict the sentiments concerning overall rating of user reviews on specific drugs. We further plan to investigate the data by training classification models among top 10 conditions and data sources. In this work we show that text mining approaches can be used to exploit similarities across the reviews of drugs related to different conditions.

2. LITERATURE

Works on drug review sentiment analysis can basically be divided into approaches applying lexicons with sentiment scores or such approaches learning sentiments employing supervised classification. In one of the earliest works on drug review sentiment analysis, a topic classifier was developed from patient data to eventually apply several polarity classifiers, one per topic [2]. Na et al. demonstrate a clause-level sentiment analysis algorithm considering multiple review aspects as overall satisfaction, effectiveness, side effects and condition. Here, a rule-based approach is employed that takes grammatical relations and semantic annotation into account and computes sentiment orientation of individual clauses based on a lexicon [4]. Gopalakrishnan et al. analyze patient drug satisfaction by using a supervised learning sentiment analysis approach. In this study three levels of polarity were classified comparing SVM with neural network-based methods [5]. Many research studies have attempted to improve domain adaption or cross-domain sentiment classification, although not on drug review aspect-level but among various entities as products, movies or restaurants. In [3] a comprehensive systematic literature review on cross-domain sentiment analysis is presented.

3. DATASET

The data used is crawled from online site Drugs.com, which is said to be the largest and most visited site by healthcare professionals and general users and was hosted on UCI Machine Learning repository[6]. The dataset provides patient reviews on specific drugs along with related conditions and a 10-star patient rating reflecting overall patient satisfaction. The data was obtained by crawling online pharmaceutical review sites. The intention was to study and perform (1) sentiment analysis of drug experience over multiple facets, i.e. sentiments learned on specific aspects such as effectiveness and side-effects,
(2) build a multi-class classifier to classify and predict condition based on reviews and evaluate the performances.
The data had 161297 reviews. Furthermore, we had features such as drugName (name of the drug), condition (name of the condition), rating (10-star patient rating), date (date of review entry), usefulCount (number of users who found the review useful). The dataset is further split into train and test sets using sklearn package. The dataset comprises of 885 Unique number of conditions and 3436 unique drugs. The average number of reviews per drug is 58.86.

4. DATA PREPROCESSING

Since the data which is imported from UCI repository is raw data directly crawled from Drugs.com, we must preprocess the data before any kind of analysis. Preprocessing the data include dealing with missing values, incorrect entries and reviews. There are null values only in the condition attribute which are about 0.05% of the data. As null values are less than 1% of the data, dropping the observations with null values will not affect our analysis. We also found 900 observations have unwanted text in the condition column instead of the actual condition. These were chosen to be dropped from the dataset. Finally, all the reviews in the data are filtered using regular expression pattern which removes all the non-alphanumeric characters.

5. EXPLORATORY ANALYSIS

We created word clouds for two features: conditions and drugName. The Fig.1 shows the most common conditions. We can observe birth control is the most common condition followed by Bipolar Disorder, Weight Loss, Emergency and others. The second one shows the most popular drug prescribed. We can map the drugs Fig.2 with the Fig.1 with conditions. For example, Ethinyl estradiol, estradiol norgestinate and etenogestral are used in birth control pills. and Venlafaxine, Gabapentin are used to treat Bipolar disorder.

Fig 1: Wordcloud of most common condition

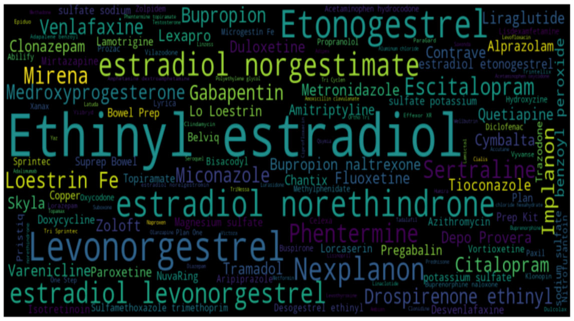

Fig 2: Wordcloud of most popular drug on market

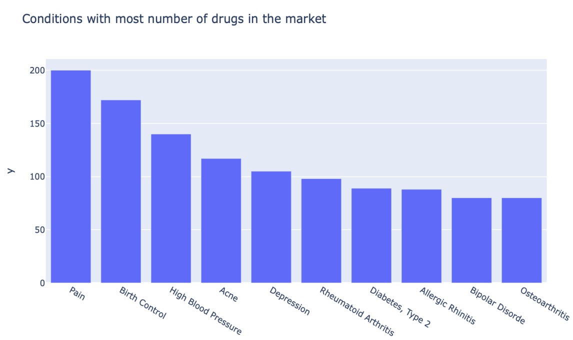

Fig 3: Conditions with highest number of drugs in market

Figure 3 shows that the highest number of drugs that are available on market are used to treat Pain followed by Birth Control, High Blood Pressure, Acne and Depression.

5.1 Rating of drugs over the years

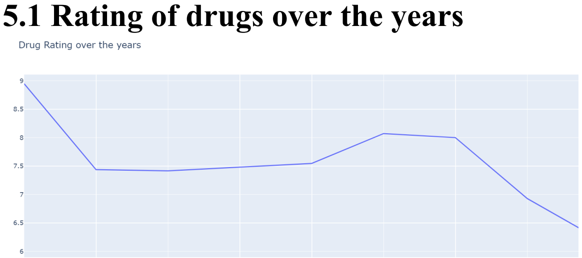

Fig 4: Ratings of drugs over the years

Drug ratings have been falling since the past 10 years. The Pharma Industry has always been an industry where people tend to give positive reviews as drugs are clinically tested on multiple levels before being sold to the public. Thus, we expect most reviews to be positive as pills are a cure for prior existing conditions .With advances in medicine and testing medicines are expected to be more reliable than ever before. However, it has been quite the contrary. We will take a look at why exactly was has this happened. Firstly 2009 was an exceptionally good year with 89% positive polarity and below 5% negative polarity.

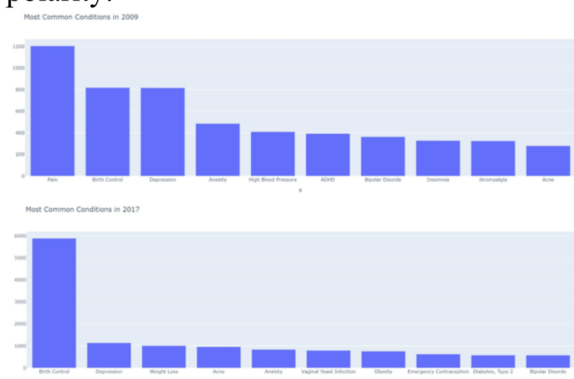

Fig 5: Most common conditions in 2009 & 2017

Let's take a look prominent conditions in 2009 and 2017.The prominent condition in 2009 is Pain while in 2017 it is Birth Control. Thus, we can come to a conclusion that abortion rates are increasing dramatically.

Fig 6: Sentiments for Pain and Birth control related drugs

Now let's take a look at sentiments associated with Pain and Birth Control related drugs. Positive sentiment associated with Pain related drug is 70.6% while negative sentiment associated with Pain related drug is 14.4%. Conversely positive sentiment associated with Birth Control related drug is 46.1% while negative sentiment associated with Pain related drug is

24.5%. So, as drugs for Birth control are sent into the market more than Pain over the years and due to the fact that those drugs are not effective the negative reviews have increased as the years progressed. This is one of the key reasons why drug ratings have dropped significantly.

6. EVALUATION METRICS

Since we are dealing with a Multi-class classification problem, accuracy is not a good metric to evaluate the performance of the model because we will not be able to discover something that is not balanced in frequency. Instead we will be using Precision, Recall and f1 score. When a confusion matrix is drawn which includes all the classes, we end up with more than one value of precision, recall and f1 scores as a result of calculations between each and every class. In order to derive a single metric, we can combine these values using Macro averaging where we compute the performance of each class and then average over classes. Given the above metrics we want to maximize precision and recall. f1 score is the harmonic mean of the two and our goal would be to maximize it. We can also look at the most important features that were learned by the model to check if the features truly belong to a certain condition which might give some insight into helping in improving the model performance.

7. MODELING AND PERFORMANCE EVALUATION

7.1 Predicting the condition based on reviews:

This task involves classifying and predicting the condition of the patient by building a multi-class classification model. The drug review dataset has a total of 885 unique conditions and 3436 unique number of drugs. If we are to use all these 885 unique drugs as categories of the target variable, the problem is going to be very complex. So, the top 10 most common conditions are filtered from the data and are used in the modeling process. The top 10 categories are 'ADHD', 'Acne', 'Anxiety', 'Bipolar Disorder', 'Birth Control', 'Depression', 'Insomnia', 'Obesity', 'Pain', 'Weight Loss'. These 10 categories serve as the classes in target variable. Data is imbalanced for the target variable. The data is split into training and testing set where 70% of data goes to training and 30% for testing. We used two most used algorithms in text classification, Naïve Bayes and Support Vector Machines. In case of Naïve Bayes, both Multinomial Naïve Bayes approach and Bernoulli Naïve Bayes approach are used. The input reviews fed to the classification model were not stemmed initially to check the performance. Then Snowball Stemmer was used to stem the reviews which showed an improvement in the performance of the model. So, stemmed reviews were used in the rest of the model building process. Stopwords were removed and all the characters are converted to lowercase. We are using the count vectorizer with term frequency for the vectorization process. The minimum document frequency is set to and ngram_range is tuned using Unigrams, Bigrams, Unigrams & Bigrams etc. were used. At the end, Hold-out test and cross-validation test results of best performing models are compared to check which method gives best results.

7.1.1 Bernoulli Naïve Bayes

ngrams	precision	recall	f1-score
(1,1)	76.15	78.05	76.63
(2,2)	56.67	74.85	58.58
(1,2)	69.59	79.01	71.88

*All values are in percentages

Table 1: Metrics of Bernoulli NB

From Table 1, we can observe that Bernoulli NB provides the best results with unigrams. When bi-grams are used there is a drastic decrease in precision which led to lower f1-score. But when both unigrams and bigrams are used, it still did perform better compared to when only bigrams were used. As it's complex to look at important features of all the 10 classes, we will look at important features learned by the best performing model of two classes: 'Birth control' and 'Obesity'.

```
-1.5868 acn          -1.7558 weigh
-1.5807 ani          -1.7426 appetit
-1.5697 spot         -1.7232 becaus
-1.5357 bleed        -1.7232 diet
-1.5145 feel         -1.7137 don
-1.5056 befor        -1.6732 tri
-1.4833 pain         -1.6551 loss
-1.4747 sinc         -1.6083 medic
-1.4286 mood         -1.5829 contrav
-1.3685 veri         -1.5527 veri
-1.3481 becaus       -1.5446 pill
-1.3334 just         -1.5366 time
-1.2904 like         -1.5340 help
-1.2870 got          -1.5234 exercis
-1.2735 time         -1.5104 lose
-1.2295 start        -1.4260 ve
-1.2268 effect       -1.3481 just
-1.2242 onli         -1.3156 onli
-1.2031 cramp        -1.2987 lbs
-1.1356 weight       -1.2476 feel
-1.0948 gain         -1.1509 pound
-1.0765 week         -1.1365 work
-1.0244 birth        -1.1049 month
-1.0075 control      -0.9094 eat
-0.9981 day          -0.8277 effect
-0.9828 year         -0.7717 week
-0.9631 ve           -0.7293 day
-0.8921 pill         -0.7210 weight
-0.5453 month        -0.6539 lost
-0.5286 period       -0.6214 start
```

From the above figures, we can see that for class 'Birth control', the model learnt features like 'bleed', 'birth', 'control' 'pill', 'period' and for class 'Obesity', we have 'weigh', 'appetit', 'diet', 'excercis', pound', 'work', 'eat'.

7.1.2 Multinomial Naïve Bayes

ngrams	precision	recall	f1-score
(1,1)	78.22	79.84	78.94
(2,2)	76.65	80.58	78.06
(1,2)	69.37	75.11	71.17

*All values are in percentages

Table 2: Metrics of Multinomial NB

In case of Multinomial Naïve Bayes, the unigrams performed much better than bigrams and combination of unigram and bigrams. For bigrams, even though the recall is slightly better than unigrams, there is a noticeable decrease in the precision which lead to a decrease in f1-score.

Most important features learned by the best performing Multinomial NB model for classes 'Birth control' and 'Obesity' are:

```
-5.1050 mood         -5.3177 tri
-5.1024 sinc         -5.3150 weigh
-5.0564 feel         -5.2966 food
-5.0391 insert       -5.2457 loss
-4.9654 bleed        -5.1808 s
-4.9568 becaus       -5.1716 veri
-4.9348 just         -5.1354 exercis
-4.9027 veri         -5.0983 time
-4.8759 like         -5.0961 help
-4.8576 pain         -5.0543 contrav
-4.8550 onli         -5.0482 lose
-4.8329 time         -5.0421 medic
-4.8275 got          -4.9457 pill
-4.8075 effect       -4.9274 onli
-4.7321 start        -4.8955 just
-4.7237 weight       -4.8560 ve
-4.7206 gain         -4.7137 work
-4.6489 cramp        -4.6793 feel
-4.6372 s            -4.6179 pound
-4.5407 birth        -4.5341 month
-4.5309 week         -4.4116 effect
-4.5064 control      -4.3419 m
-4.4849 year         -4.2967 lost
-4.4455 m            -4.2032 eat
-4.3738 day          -4.1663 t
-4.3428 ve           -4.1129 week
-4.1729 pill         -4.0939 lbs
-3.9081 t            -4.0714 weight
-3.8513 month        -4.0242 start
-3.8160 period       -4.0054 day
```

The features learned for 'Birth control' include 'bleed', 'birth', 'control', 'pill', 'period' and for 'Obesity' the features are 'weigh', 'food', 'loss', 'lbs', 'excercis', 'start'.

7.1.3 Support Vector Machine

ngrams	precision	recall	f1-score
(1,1)	74.05	74.83	74.41
(2,2)	63.99	65.15	64.43
(1,2)	74.53	75.60	75.00

*All values are in percentages

Table 3: Metrics of SVM

Unlike Naïve Bayes, where unigrams performed better, For Support Vector Machine, combination of unigrams and bigrams gave better performance. The performance of model using unigrams is very close (< 0.5%) to the unigram and bigram model. Bigrams gave the worst results for SVM though it did better in case of Multinomial NB.

Most important features learned by the best performing SVM model for classes 'Birth control' and 'Obesity' are:

```
0.6711  nuvar
0.6779  best birth
0.6845  love recommend
0.6910  noth good
0.6933  sleepi s
0.6945  caus gain
0.6969  year effect
0.7042  effect s
0.7291  say noth
0.7407  acn total
0.7434  total gone
0.7470  lutera
0.7485  avian
0.7736  stroke
0.7781  implanon
0.7834  just love
0.7876  complaint
0.7974  problem feel
0.8328  realli issu
0.8620  damn
0.8625  skyla
0.8688  great issu
0.8851  loss weight
0.9200  ani problem
0.9209  intend
0.9468  gain tired
0.9757  mirena
0.9988  pregnant
1.1258  insert
1.1612  great skin
```

```
0.6173  alway hurt
0.6205  great say
0.6236  taken phentermi
0.6291  work tri
0.6414  stomach alway
0.6452  drug weight
0.6484  diethylpropion
0.6497  phendimetrazin
0.6527  great stop
0.6575  problem lost
0.6652  hcg
0.7172  medic march
0.7286  meridia
0.7311  great alreadi
0.7316  great lost
0.7542  lost lot
0.7556  bontril
0.7602  experi effect
0.7702  orlistat
0.7703  just work
0.7892  hi start
0.8234  just night
0.8942  day seen
0.9183  xenic
0.9803  work increas
1.0035  alli
1.0180  suffer year
1.0282  readi tri
1.1786  experi good
1.3286  tenuat
```

The features learned for ‘Birth control’ are ‘best birth’, ‘just love’, ‘complaint’, ‘pregnant’, ‘loss weight’ and features learned for ‘Obesity’ are ‘work tri’, ‘diethylpropion’, ‘lost lot’, ‘experi good’.

7.1.4 Summarizing metrics of best models

	ngrams	precision	recall	f1-score
Bernoulli NB	(1,1)	76.15	78.05	76.63
Multinomial NB	(1,1)	78.22	79.84	78.94
SVM	(1,2)	74.53	75.60	75.00

*All values are in percentages

Table 4: Summary of best models

From the above table, it is clear that Multinomial Naïve Bayes with unigrams gives the best performance among all the other variations of parameters and algorithms.

7.1.5 Hold-out Method vs Cross Validation

We also implemented cross-validation for best performing models of Bernoulli and Multinomial NB and Support Vector Machine.

For cross validation scores, 10-fold cross validation was performed using the best models to compare the performances.

model	hold-out score	cross-val score (10-fold)
Bernoulli NB	82.45	79.93
Multinomial NB	84.30	84.25
SVM	81.59	81.76

*All values are in percentages

Table 5: Hold-out vs Cross Validation scores

7.2 Sentiment Analysis

So, what exactly is a sentiment? Sentiment relates to the meaning of a word or sequence of words and is usually associated with an opinion or emotion. For analyzing patient reviews it was imperative on our part to perform sentiment analysis to retrieve the polarity of emotion generated from the reviews given by patients .With Sentiment analysis our aim was to mine patient reviews to identify and extract subjective information in patient reviews and help build a generalized understanding the effectiveness of the drugs. Keeping this in mind we used two algorithms Naïve Bayes and Support Vector Machine to classify data .We used the hold out test method where 75% of data was training data and the rest 25% was used for testing. Our focus is to maximize recall by minimizing the number of false negatives. To ensure this we tuned vectorizers and played with a few combinations to retrieve the best results

While Choosing a vectorizer we had four options

1] Boolean : Tells if a word is present or absent in a document

2] Term Frequency : Counts the word frequency

3] Normalized Term Frequency : Term frequency divided by the total word count of that document – necessary for documents of different lengths

4] TFIDF : Term frequency just gives the occurrences of a word in a document. If that word occurs in most of the documents irrespective of the target variable categories, it will not be useful for text categorization. IDF Penalizes words occurring in most documents as these words might be stopwords or topic or title words that won't add true meaning.

After Hyper-parameter tuning, we get the best generalized model with the following parameters

- strip accent='ascii' to remove all non-English tokens from occurring in feature attribute.
- stop_words='english' Stop words are removed from the text because they add no true meaning to the corpus.
- min_df=3 a word needs to occur more than 3 times. i.e. cutoff frequency should be above 5 for word to be selected in the features attribute
- lowercase=True

7.2.1 Naïve Bayes Model

Bernoulli Naïve Bayes, a document is a binary vector over the space of words. The dimension of the vector for the document is either a 1 or a 0. This indicates whether a word occurs at least once in the document. The probability of a label given its class is simply the product of the probability of the attribute values over all word attributes.

Bernoulli NB(Boolean,alpha=0.01)				
Accuracy	0.94	0.59	0.79	
Precision	0.80	0.81	0.86	
Recall	0.79	0.59	0.94	0.84
Bernoulli NB(TF,alpha=0.01)				
Accuracy	0.92	0.58	0.78	
Precision	0.79	0.76	0.86	
Recall	0.78	0.58	0.92	0.83

Table 6: Results of Bernoulli Naïve Bayes for Sentiment Analysis

Bernoulli Naïve Bayes with Boolean vectorizer with using ngram_range (1,3) i.e. combination of unigrams, bigrams and trigrams is the best performing model.

Multinomial Naïve Bayes model captures word frequency information in documents. It is assumed that the lengths documents are independent of class. This model uses a bag of words representation of the documents where the number of times a word occurs in a document is counted. Thus, the probability of a document given its class is simply the multinomial distribution. Because of the independence assumption, the parameters for each attribute can be learned separately. This results in a sparse matrix problem which can be resolved through smoothening to avoid posterior probabilities from becoming zero.

Multinomial NB(Boolean)				
Accuracy	0.92	0.60	0.74	
Precision	0.82	0.70	0.86	
Recall	0.74	0.60	0.92	0.82
Multinomial NB(TF)				
Accuracy	0.92	0.60	0.74	
Precision	0.82	0.69	0.85	
Recall	0.73	0.59	0.91	0.83
Multinomial NB(TF,without ngrams,Lemmatizer)				
Accuracy	0.81	0.37	0.57	
Precision	0.61	0.39	0.78	
Recall	0.57	0.37	0.81	0.68
Multinomial NB(TF,without ngrams,without Lemmatizer)				
Accuracy	0.81	0.41	0.61	
Precision	0.65	0.40	0.80	
Recall	0.61	0.41	0.81	0.69
Multinomial NB(TF,with nram ,without lemmatizer)				
Accuracy	0.92	0.60	0.74	
Precision	0.82	0.69	0.85	
Recall	0.73	0.59	0.91	0.83

Table 7: Results of Multinomial Naïve Bayes for Sentiment Analysis

We get best results with TF-Vectorizer for Multinomial NB. After performing hyper-parameter tuning (0,0.01,0.5,1)of smoothing parameter we get value of alpha as 0.01 for both Bernoulli and Multinomial NB which is added to word count in the given class. This smoothing regularizes Naive Bayes and since we don't add 1, this type of smoothing is called Lidstone smoothing**.**

7.2.2 Linear Support Vector Machine

SVM draws that hyperplane by transforming our data with the help of mathematical functions called kernels. Since the problem is linearly separable as we want to determine whether the drug will be effective or not, we make use of Linear SVC. After performing GridSearchCV for Cost Parameter (0,0.01,0.5,1,5,10). We get value of slack variable as 1 which is the default value as it tries to maintain a bias-variance trade off by ensuring misclassification of examples is penalized but not a high degree as it may result in overfitting.

Linear-SVC(lemmatizer,tfidf),C=1				
Accuracy	0.88	0.24	0.63	
Precision	0.62	0.48	0.77	
Recall	0.63	0.24	0.88	0.71
Linear-SVC(no lemmatizer,tfidf),C=1				
Accuracy	0.91	0.28	0.69	
Precision	0.69	0.55	0.80	
Recall	0.69	0.28	0.91	0.75
Linear-SVC(no lemmatizer,tf, with ngram(1,3)),C=1				
Accuracy	0.96	0.67	0.86	
Precision	0.86	0.84	0.90	
Recall	0.86	0.67	0.96	0.89

Table 8: Results of SVM for Sentiment Analysis

So, Linear SVC with term frequency and combination of unigrams, bigrams, trigrams and C=1 is the best performing model with precision of 86% and recall of 86%.

After building and evaluating the models constructed on Naïve Bayes and Support Vector Machine, we see that using a lemmatizer reduces the accuracy of the model. Also adding bigrams and trigrams increases accuracy considerably. From all the above tables we can observe that Linear SVC with term frequency input and ngram range of (1,3) with precision of 86% and recall of 86% beating best performing Bernoulli NB and Multinomial NB by almost 6% and 4% in terms of precision and by 7% and 12% in terms of recall respectively, is the best performing model for Sentiment Analysis.

Most important features learned by the best performing model:

```
(-11.288673754857625, 'love')
(-9.996507072184688, 'amazing')
(-9.535321924939172, 'miracle')
(-8.602134252823786, 'changed life')
(-8.045122937143175, 'saved')
(-7.503849727001853, 'best')
(-7.2714085097883885, 'wonderful')
(-6.598205925051163, 'excellent')
(-6.360135343914587, 'years')
(-6.236145826641111, 'great')
(-6.160443964575975, 'works')
(-5.885334614086755, 'wonders')
(-5.881765029081379, 'highly recommend')
(-5.622501792984056, 'fantastic')
(-5.531428800493868, 'cleared')
(-5.515645832112049, 'gone')
(-5.393933554977023, 'awesome')
(-5.286933082884927, 'thank')
(-5.217941469806139, 'lifesaver')
(-5.201289170230897, 'gave life')
```

Positive features

```
(8.55364430491418, 'quot gentle')
(8.55364430491418, 'strain relationship')
(8.55364430491418, 'taking poison')
(8.607682302226653, 'calling doc')
(8.607682302226653, 'swings worst')
(8.658949294321957, 'nexplanon removed')
(8.658949294321957, 've thing')
(8.658949294321957, 'wait switch')
(8.658949294321957, 'worse decision')
(8.707715660583224, 'drawing board')
(8.754214041264685, 'money product')
(8.754214041264685, 'rebound flushing')
(8.754214041264685, 'respiclick')
(8.754214041264685, 'stopping immediately')
(8.79864604977359, 'horrible product')
(8.881992890615843, 'bleeding worst')
(8.881992890615843, 'crotch')
(9.273905292124379, 'refund')
(9.300566428727473, 'horrible medication')
(9.469601833031156, 'awful drug')
```

Negative features

Some of the most positive features learned by the model are 'amazing', 'miracle', 'excellent', 'fantastic', 'lifesaver' whereas negative features are 'strain relationship', 'swings worst', 'worse decision', 'horrible product' etc.

7.2.3 Error Analysis

Positive sentiments predicted as Negative:

"Works well!! Second time taking it and I'm getting all but 4 side effects. Nausea is brutal. Lightheaded/dizzy. Abdominal cramps. IBS flare! Moody. Headache, calf pain.... etc....etc. Stopped today after calling the pharmacist. Side effects are just too severe. No problem the first time I used it. And nooooo booze when taking this! You'll get violently ill. Seriously."

->sentiments associated with side effects

"Works right like it said it would but heavy bleeding

->sentiments associated with side effects

Linguistic Patterns

Negative emotions associated with Side Effects overpower the real meaning of the sentence

Negative sentiments predicted as positive:

"Loved this medicine, helped me a lot. Problem is I got glaucoma"

->Interesting Linguistic Pattern noted is that Pill worked and helped patient, but next sentence implicitly states he still has a problem and wrong medicine was prescribed.

'Was originally on Oxy-neo and later Hydromorphone which worked but no longer wants to take opiates long-term and am looking for other solutions. Unfortunately, Gabapentin is not working"

->Pill Working but patient does not want the pill due to fear of aftereffects which are implicit.

"7.5 mg dosage has done nothing for me. I think I should be trying the max dose of 15mg as I had taken the max dose 4mg of Detrol LA"

->Interesting Linguistic Pattern noted is the fear of taking medicine in huge amounts which explains about the ineffective composition of the medicine

Linguistic Pattern

Lot of underlying implicit meanings like fear which SVM was not able to classify correctly

8. Conclusion & Future work

From the results it is found that Multinomial Naïve Bayes is the performing model for condition prediction with precision of 78.22% and recall of 79.84%. For Sentiment Analysis, Linear SVC gave us the best results with precision of 86% and recall of 86%. Within this project, we studied the application of machine learning algorithms on text data in supervised classification and sentiment analysis of patient generated drug reviews. Besides patient condition prediction, sentiment aspects concerning effectiveness of the drugs were analyzed. We learnt how to work with unstructured data that is obtained from the real-world applications. This has also helped us learn how to process the raw data and find hidden insights from it.

However, there is some room for improving the performance of the models. The results could be increased further by eliminating the some more errors like misspelling of words. Although these experiments are carried out on a large dataset and imbalanced dataset, there might be some factors which questions the validity of the results. To test further, we can use datasets obtained by various collecting methods and of different sizes. Also using deep learning approaches opens to numerous possibilities of new insights that can be obtained from the reviews. This will help in leading to a better understanding of patient reviews.

References

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). Association for Computing Machinery, New York, NY, USA, 121–125.*
2. *Lei Xia, Anna Lisa Gentile, James Munro, and José Iria. 2009. Improving Patient Opinion Mining Through Multi-Step Classification. In Proceedings of the 12th International Conference on Text, Speech and Dialogue (TSD '09). Springer-Verlag, Berlin, Heidelberg, 70–76.*
3. *T. Al-Moslmi, N. Omar, S. Abdullah, and M. Albared. 2017. Approaches to Cross- Domain Sentiment Analysis: A Systematic Literature Review. IEEE Access 5 (2017), 16173–16192.*
4. *Jin-Cheon Na and Wai Yan Min Kyaing. 2015. Sentiment Analysis of User- Generated Content on Drug Review. Journal of Information Science Theory and Practice 1, 1 (Mar2015).*
5. *Vinodhini Gopalakrishnan and Chandrasekaran Ramaswamy. 2017. Patient opinion mining to analyze drugs satisfaction using supervised learning. Journal of Applied Research and Technology 15, 4 (2017), 311 – 319.*
6. https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29