Topic Modeling of Drug Reviews

Chiau Yin Yang
IST736 Text Mining - Syracuse University
Syracuse, NY, 13210
cyang38@syr.edu

Abstract

With the rapid growing usage of online customer reviews, it is vital for business to understand its customers' feedback quickly and accurately. When it comes to text mining, sentiment analysis, e.g. positive or negative of a review, and classification of such analysis on customer reviews are generally popular in this area of study. In this study, the online drug reviews from drugs.com were analyzed. Rather than sentiment analysis, useful count on a patient's review was used as target attribute using machine learning technique - Naive Bayes. Additionally, topic modeling is used to obtain overview of what patients wrote about the products.

I. Introduction

This research is to analyze the patient reviews on 2,637 different drugs that focus on 708 associated conditions. Patients left the reviews with ratings that are on a scale of 1 to 10 based on patients' overall satisfaction.

The primary objectives of this research is to find out what topics were these reviews mentioning and whether machine can discern and predict the usefulness of a review. Topic modeling (Latent Dirichlet allocation, LDA), a classification models - Multinomial Naive Bayes, and other text mining techniques were performed to achieve the goals.

This data set was obtained from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29), and it was collected from online pharmaceutical review sites - drugs.com - on 10/04/2018.

I. Data Set Overview

Originally, the .tsv file has 53,766 instances and 6 attributes, which are as below screenshot. The left screenshot shows the all attributes from original data set, and the right one shows the final view. To better analyze the reviews, a few pre-processing steps and other variables were added.

```
drugName        object
condition       object
review          object
rating         float64
date            object
usefulCount      int64
dtype: object
```

```
drugName                object
condition               object
review                  object
rating                 float64
date            datetime64[ns]
usefulCount              int64
Month                    int64
Year                     int64
period                 float64
useful_ct_by_m         float64
useful_dummy             int64
length                   int64
dtype: object
```

- Drop duplicate rows - there were no duplicates or NAs.
- "Useful count by month" is calculated by "useful count" dividing "period"
- Convert "date" column into datetime component and broke down to "month", "year" and "period (by month)". The period is calculated by subtracting reference date (10/04/2018) to the date.
- "Useful dummy" is a labelled-dummy variable that was coded by the quartile of useful count. 0 being less useful and the higher the number the more useful count it has.
- "Length" represents the word count of each review after the text was preprocessed with Python Scikit learn CountVectorizer. Stop words

were not removed when performing this task.

II. Descriptive Analysis

To further understand the dataset, I summed up all the useful counts and number of unique reviews by conditions as it is a more understandable categorical variable but not too many unique groups like drug name. Based on below results, the number of reviews represents the popularity of a specific condition. This could indirectly indicate that the top 5 issues most patients are looking to fix are: birth control, depression, pain, anxiety and acne. In terms of useful count, it represents these reviews provided useful information for related conditions. In other words, the number of review indicates the number of sales of a specific symptoms, whereas the number of useful count could indirectly influence the purchase decision for future patients. According to below results, reviews for depression are more than twice as many as reviews for birth control, although there were three times more reviews about birth control than depression.

There are various reasons why this happened, maybe patients who have depressions are less willing to leave reviews, or those who left reviews for birth control did not provide enough useful information or had diverse opinions.

	condition	review
122	Birth Control	8402
201	Depression	2798
478	Pain	1898
87	Anxiety	1748
49	Acne	1694
121	Bipolar Disorde	1236
676	Weight Loss	1138
347	Insomnia	1127
447	Obesity	1055
44	ADHD	980

	condition	usefulCount
201	Depression	163185
87	Anxiety	93409
122	Birth Control	77356
478	Pain	74249
676	Weight Loss	49163
121	Bipolar Disorde	48275
447	Obesity	44788
347	Insomnia	42777
44	ADHD	39382
304	High Blood Pressure	38270

III. Topic Modeling

There are various algorithms to performing topic modeling, and the one I used in this research is Latent Dirichlet allocation LDA).

A. Data Preprocessing

CountVectorizer and TF-IDF vectorizer function in Scikit-learn python package are used to tokenize the reviews with document frequency ranging from minimal document frequency = 5 to 95% of the documents. The max feature is 1,000, and basic built-in stopword list were used to clean the text. I started with 4 topics, which the number is selected randomly. As you can see below, there are terms that do not provide much information, such as “039”, or “ve”.

Without changing other parameters, I updated the stopword list by adding a few more words, including “039”, “ve”, “don” etc., to see if the results could be more informative. Here is the result after preliminary stopword removal.

	Top 30 number of words	Label
0	day days took pain like just time taking hours bad skin did use went effects doctor night using started felt got stomach work prescribed used medicine medication nausea minutes worked	About specific side effects or results
1	years pain anxiety taking life day effects feel sleep like medication mg depression medicine time drug doctor started just night quot better work months tried felt works dose great weeks	About depression/anxiety
2	weight started lost taking months week effects weeks	About specific

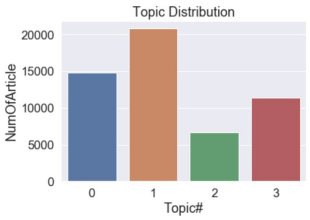
	pounds years day eat month just loss appetite lbs year feel gain days gained ago eating time good far old lose great	effects/results about weight loss, appetite
3	period pill control months birth got month cramps periods sex days acne bleeding just time like mood years week swings weeks bad started weight having spotting getting drive didn effects	About birth control and acne/skin

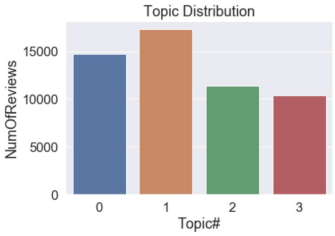
Next, to better label and interpret those topics, I focus on nouns that are related to symptoms and effects, and adjectives that describing feelings and experience, so I further removed less meaningful verbs like “take”, “go”, “do”, “get”, “feel”, and time adverbs like “year”, “month”, “time”, “day”, and see if more interesting results can be observed. I also contemplated whether including more words (number of feature changes from 1,000 to 10,000) can improve the results, so I performed a few more experiments.

Below result could be the best among all results even though you can see multiple symptoms within one topic.

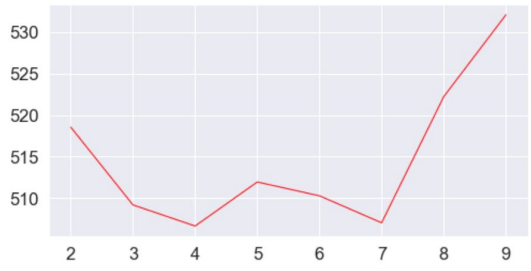
	Top 30 number of words	**Label**
0	pill period control birth cramps periods sex bleeding mood like started having taking bad love really acne pills effects haven normal experience little horrible body recommend went stopped experienced ago	Birth control (sex)
1	pain weight skin far gain started	Weig
	acne lost taking effects severe good lot bad really ago little headaches went doctor help noticed better like using away medicine right worse thing	ht loss, acne/skin, headache (pain reliever)
2	anxiety started quot life like night sleep medicine effects depression mg medication doctor feeling dose better ago morning went help prescribed work able helped bad having really finally good	Depression, anxiety
3	works drug effects great medication use worked used work stop like tried using doctor taking really pills better gave try right symptoms know good make long body bad think went	Cannot distinguish

Based on above result, you can see that the topics are differentiated by various symptoms (conditions). For the result of 4 topics, topic distribution becomes evenly after removing some non-meaningful verbs and times (month, year, day).





To verify if the random number of topics above is the best number for this practice, I ran perplexity measurement for range between 2 and 9. Based on the result, 4 and 7 are the best number of topics for these reviews. Since the first random number I chose was 4, I performed another LDA topic modeling with 7 topics. Yet, the results did not help distinguish topics from topics significantly, compared to the result of 4 topics, because there are more similar topics.



0	started taking weeks week ago went use time doctor using like lost better away old effects stopped right life prescribed having work try used good little really lot help symptoms	
1	medicine taking sleep hours night mg works dose effects great time prescribed doctor nausea bad work drug good worked able effect later make long does try best started like know	Side effect (nausea)
2	medication morning effects like work tried think worked headaches used really stop night time does helped body effect having know want great prescribed little severe experienced able experience normal long	experience
3	pain skin acne severe help bad worse helped doctor away tried better recommend really right great went used lot finally say does little worked like experience long time able try	Experience related to pain
4	anxiety quot life depression feeling drug like effects better symptoms worse noticed horrible normal help severe helped bad tried weeks time doctor finally really having mood started say stop mg	Anxiety, depression, dose
5	pill period control birth cramps periods sex bleeding time taking haven week acne mood started having weeks bad horrible experience effects normal really experienced headaches body stopped stop recommend noticed	Birth control
6	weight gain pills good thing gave like far great love effects really bad best little say control lot think works try make tried mood recommend better old want birth does	Mix of symptoms

B. Topic interpretation

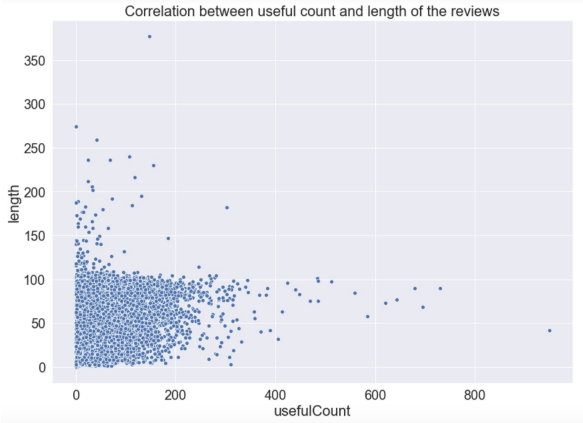
All in all, the result generated from LDA topic modeling mostly matches the top 10 conditions by number of reviews. The information a word can bring is slightly vaguer than a few more words but you can still perceive that the reviews touch on different experience, feeling or side effects with certain symptoms.

At the first glance, the top 30 words of each topics contain terms that could involve different conditions, making topic labeling slightly more difficult and not as straightforward as expected. However, some drugs have side effects that are similar to

certain conditions. For example, some birth control drugs could bring about skin breakout, weight-loss or gain, bleeding so on and so forth. Thus, it does make sense to see mixture of different conditions (terms). Furthermore, some symptoms, birth control and depression/anxiety, are more distinct that others.

IV. Predictive Analysis

In this section, I set a hypothesis that the machine can tell if a review is useful or not. Before moving onto the predictive algorithms, I would like to know if there is any correlation between the length of the review and useful counts. From below scatter plot, there is no clear trend that the longer the review, the more information people find useful. Most of the reviews are within 100 words and approximate 0 to 200 useful counts.



The model I used in this research is Multinomial Naive Bayes as there are 3 or 4 target attributes. Multinomial Naive Bayes is part of the Naive Bayes algorithm, which is a probabilistic classifier, and is based on Bayes' theorem with a naive assumption that the features are independent from one another. The Bayesian theory formula is as follows:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

In the modeling, I used held-out test split, 60% training and 40% test, and tried different parameters to observe the changes. In the first experiment, I divided the target attribute into 3 classes. The distribution for this classification is uneven because there are more class "useful" (target label = 1) than other classes.

- 0: Less useful (useful counts are < 25% quartile);
- 1: Useful (useful counts are between 25% and 75%);
- 2: More useful (useful counts are >= 75%)

Based on the result, TF-IDF tokenizer has slightly better accuracy than Count Vectorizer. Uni-gram alone has better performance than bi-gram and combined.

However, the time component may influence this variable because the older the reviews, the higher possibility of having more useful counts. Thus, in the second experiment, I normalized the number of useful counts by dividing the "period" (month). Furthermore, instead of 3 classes, 4-class is used for the purpose of even distribution.

- 0: Less useful (useful counts per month are < 25% quartile);
- 1: useful (useful counts per month are between 25% and 50%);
- 2: More useful (useful counts per month are between 50% and 75%)
- 3: Most useful (useful counts per month are >= 75%)

Based on the result, the accuracy rate has dropped tremendously from the first experiment. This could be because the algorithm is not able to distinguish detailed

groups, meaning there is no obvious difference between the "useful" group (2) and the "more useful" group (3), or because there are more "less useful" group (1) in the first experiment.

Without knowing the mechanism of the review website, I do not know the degree of the "period" variable affecting the result because using "useful count per month" implies website visitors read all historical reviews equally while browsing. The way those reviews display could be chronological or by the number of useful counts. Therefore, I conducted the third experiment that uses number of useful count instead of the one by month.

The accuracy does improve a little though it is not the best among the 3 experiments. However, since this experiment has balanced distribution on the target attribute, this is the best result by far.

Below is the summary of all three experiment:

- First experiment (3 class - uneven distribution)

	CountVector	TF-IDF
Uni-gram	0.546938	0.566280
Bi-gram	0.539405	0.561026
Uni + Bi-gram	0.520388	0.538801

- Second experiment (4 class - even distribution; normalized useful count by month)

	CountVector	TF-IDF
Uni-gram	0.417398	0.412377
Bi-gram	0.425349	0.417305
Uni + Bi-gram	0.374203	0.377644

- Third experiment (4 classes - even distribution; useful count)

	CountVector	TF-IDF
Uni-gram	0.43753	0.439484
Bi-gram	0.43181	0.440042
Uni + Bi-gram	0.39470	0.397358

V. Results and interpretation

Since the third experiment is the best result, I will interpret this result. Based on below screenshot, you can see that for class 0 and 3, it has higher accuracy rate compared to class 1 and 2. It means the computer cannot distinguish the nuance between "useful" and "more useful" groups.

```
[[2785  899  624  648]
 [1717 1328 1115 1331]
 [ 643 1027 1621 2326]
 [ 227  480  981 3755]]

The accuracy score is 0.44120518900822986
The summary is
              precision    recall  f1-score   support

           0       0.52      0.56      0.54      4956
           1       0.36      0.24      0.29      5491
           2       0.37      0.29      0.33      5617
           3       0.47      0.69      0.56      5443

   micro avg       0.44      0.44      0.44     21507
   macro avg       0.43      0.45      0.43     21507
weighted avg       0.43      0.44      0.42     21507
```

Within the same settings, I changed the max feature from "None" to 300 and 500, and the accuracy did not increase at all. Moreover, I compared the polarized target groups using bi-gram TF-IDF vectorizer setting. Before updating stopword list, although the accuracy is higher, it is more understandable for human. There were 8 identical pairs of bi-grams, and 5 identical pairs of bi-grams after removing certain stop words between the "less useful" (class =0) and "most useful" (class =3). Therefore, the result shows it is easier to tell the difference (more different words) between two groups.

Less useful group (target attribute = 0)	
Before updating stopwords	After updating stopwords

took pill')	lose weight')
works great')	read reviews')
went away')	highly recommend')
years ago')	cystic acne')
stopped taking')	breakthrough bleeding')
gain weight')	heavy bleeding')
yeast infection')	unprotected sex')
got period')	works great')
really bad')	went away')
years old')	stopped taking')
months ago')	gain weight')
felt like')	yeast infection')
gained weight')	really bad')
feel like')	gained weight')
started taking')	started taking')
taking pill')	taking pill')
sex drive')	sex drive')
weight gain')	weight gain')
mood swings')	mood swings')
birth control')]	birth control')]

Most useful group (target attribute = 3)	
Before updating stopwords (0.44)	After updating stopwords (0.43)
depression anxiety')	stopped taking')
weight loss')	anxiety panic')
months ago')	life saver')
twice day')	taking medication')
year old')	chronic pain')
doctor prescribed')	hot flashes')
feel better')	sleep night')
felt like')	anxiety depression')
times day')	highly recommend')
years old')	weight loss')
dry mouth')	depression anxiety')
weight gain')	doctor prescribed')
saved life')	weight gain')
changed life')	dry mouth')
works great')	saved life')
years ago')	changed life')
feel like')	works great')
started taking')	started taking')
panic attacks')	panic attacks')
blood pressure')]	blood pressure')]

However, in the two middle classes, there are 14 duplicate pairs in top 20 bi-grams, so it is harder to discern the two groups.

As for the stop words list, after adding non-meaning verbs and some time adverbs, it is more discernible to interpret than before.

VI. Conclusion

In summary, the result generated from topic modeling aligns with the top 10 conditions by number of reviews. Although the accuracy rate for the Naive Bayes is not the best (0.44), you can see that the algorithm performs better at identifying the polarized groups (0 and 3) but not the middle groups (1 and 2). Furthermore, the length of the reviews may not have pronounced correlation to the number of useful counts.

VII. Future work

Due to the time limit, this research partially serves the purpose of experimental analysis, so not all options were fully explored. Here are the future works to make it better: (1) Stop word list should still be updated and tailored to this specific

research. (2) Stemmering can be performed to combine same words with different tense and singular/plural words. (3) Sentiment analysis based on ratings can be performed and compared. (4) More studies are needed on how useful count works before classifying the attribute. (5) In topic modeling, can examine specific word, like “experience”, “side effect” to compare results.

VIII. Reference

[1] Blei, D. et al., (2003) Latent dirichlet allocation. *Journal of machine learning research* 3(4-5):993-1022
[2] Wikipeidia Naive Bayes, Bayes theorem. https://en.wikipedia.org/wiki/Naive_Bayes_classifier#Multinomial_naive_Bayes
[3] UCI Machine Learning Repository