Discover Customers' Health Care Demand – Drug Review Analysis by Natural Language Processing

Xinyi(Sarah) Zhao
xinyiz@umich.edu

Abstract

Understanding patients' demand on medication products and their perceptions about the products performance and related services are important for pharmaceutical industries to succeed in the competitive market. Most research on medication products analysis is based on consumers' purchase history and drug reviews, the regular analysis on product reviews are mainly focused on sentiment analysis about the positive and negative comments. However, the more specific attributes about the performance of the drugs, for example "effectiveness", needs to be explored and there are lots of details about patients' needs that cannot be discovered by this approach. In this study, we retrieved the drug reviews and related information from WebMD, a widely used website about medication, and built a prediction system to uncover the performance of drugs from patients reviews by natural language processing models and investigated the patients' needs by text-mining. We found that some important attributes of drugs can be estimated by deep learning approaches and the key elements about people's demand can be explored by topic modeling and related NLP techniques. This kind of study would be beneficial to the pharmaceutical industry on market forecast and new products design.

Introduction

Since the COVID-19 pandemic in 2019, disease prevention and health improvement play more and more important roles in people's life. With the rapid development of the pharmaceutical industry, large numbers of new medications have been developed and launched to market. Although the medication products were approved by FDA, it doesn't mean the products are effective for each patient. Thus, it becomes very critical to understand the performance and related features of the drugs on users. Recent years, text and natural language analysis captures more attention as it can explore deeper information and bring more insight into the quantitative methodology, so it can be used to analyze users' demand.

The previous pharmaceutical market research and prediction are mostly based on sales record and patients' order history, but this kind of information is hard to collect and a large part of this data is confidential. In the meantime, we found that drug reviews from patients is a good resource to be used for related analysis and prediction. In this study, I did some drug review exploration by nlp techniques and found the key words or phrases that are important to represent the health demand and critical health needs of patients, and also built a prediction system to predict the effectiveness, easy-to-use and satisfaction of the drugs.

This kind of study will be helpful for the pharmaceutical industry to better understand consumers' demand and health needs, as well as patients' responses for corresponding medications. The marketing managers or business leads will also care about these studies because understanding consumers' needs, or intentions will be helpful for the marketing forecast. In addition, improving the effectiveness and ease-to-use of drugs will benefit patients' health and make their life healthier.

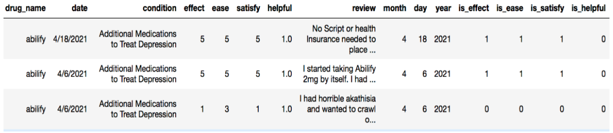
Data

The main data will be collected form WebMD drugs reviews (https://www.webmd.com/drugs/2/index) by web scraping. The drug reviews include the drug review text, four numerical variables, "Effectiveness", "Easy to Use", "Satisfaction" in the scale of 1 to 5 and number of votes for "helpfulness" (the score starts from 0). It also includes the date of posting, the related information about the patients and the patients' conditions.

The data include more than 59559 instances after dropping the null values, which include the most recent reviews. The median score for "effectiveness", "ease to use" and "satisfaction" are 4.0, 5.0 and 3.0 respectively and the median count of “helpfulness” is 6.0 (the value of helpfulness is in the range of 1 to 158).

The review text includes some noise, eg. “<br/>” or some digital strings. The noise has been removed by regular expression and the stop words have been removed while tokenizing. Finally, we got totally 1,887,270 tokens and 40776 unique tokens.

By count the frequency of each unique tokens, we have got the top 30 most common used words from the review text, which include “pain”, “take”, “side”, “effects”, “day”, “needed”, “drug”, “insurance” and “order” etc. This result gives us some hints about the demand of patients on drugs.



drug_name	date	condition	effect	ease	satisfy	helpful	review	month	day	year	is_effect	is_ease	is_satisfy	is_helpful
abilify	4/18/2021	Additional Medications to Treat Depression	5	5	5	1.0	No Script or health Insurance needed to place ...	4	18	2021	1	1	1	0
abilify	4/6/2021	Additional Medications to Treat Depression	5	5	5	1.0	I started taking Abilify 2mg by itself. I had ...	4	6	2021	1	1	1	0
abilify	4/6/2021	Additional Medications to Treat Depression	1	3	1	1.0	I had horrible akathisia and wanted to crawl o...	4	6	2021	0	0	0	0

Related Work

Regular studies on product reviews mainly talk about sentiment analysis and classification, for example, the article “Sentiment analysis using product review data" by Xing Fang et.al. discussed the typical sentiment mining to classify the reviews to positive and negative[1], and the example “Towards Enhanced Opinion Classification using NLP Techniques" by Akshat Bakliwal, talks about the algorithms that classify the reviews by the positive or negative opinions on the products[2]. Another study by Shivaprasad T. K. et.al, used both machine learning approaches and Lexicon based approach for the classification[8]. Additionally, the article “Customer Perception Analysis Using Deep Learning and NLP" talked about uncovering consumers’ interests by Contextual Semantic Tagging and deep learning[9]. And the report “Very quaffable and great fun: Applying NLP to wine reviews" built a prediction system with Word2Vec and LDA[10]. In this study, I have not considered the reviews simply by positive or negative based on the rate, I built a prediction system to find the hidden information of the reviews, such as the “effectiveness”, “ease-to-use”, “helpfulness” and “satisfaction”, in order to understand users’ demand in this special period. In addition, the data exploration by tokens, bigrams and trigrams are also informative for uncovering patients’ needs, which has been emphasized by some report, eg. “Towards Enhanced Opinion Classification using NLP Techniques" by Akshat Bakliwal et.al[4].

Furthermore, text-mining and topic modeling is also an important aspect for review analysis, this technique has been developed in recent years. This paper “Text-mining of PubMed abstracts by natural language processing to create a public knowledge base on molecular mechanisms of bacterial enteropathogens"[5] provided us with good suggestions. There are tons of studies talk about topic modeling and related measurement, for example “Measuring service quality from unstructured data: A topic modeling application on airline passengers’ online reviews" by Nikolaos Korfiatis gives us good example on measuring quality by reviews[6]; and “Automated topic modeling of tourist reviews: Does the Anna Karenina principle apply?" by Andrei P. Kirilenko talks about the common problems and solutions of topic modeling[7]. In my study, I also tried the topic modeling to extract topics and keywords from the drug reviews, and furthermore, I built a prediction model with features created by topics keywords and compared it with my first prediction system.

Methods

1. Data collection by web scraping, the data come from the WebMD website (https://www.webmd.com/), because the html file of the websites are blocked by “Cloudflare”, I used “cfscrape” and “bs4” packages for web scraping and collected more than 70 thousand instances.

2. Data clean and organizing, the original review text contains lots of noise and null values, which has been removed, then a dataframe with 14 columns has been created (include “date”, “condition”, “effect’, “ease”, “satisfy” and “helpful” etc), the “is_effect”, “is_ease”, “is_satisfy” and “is_helpful” are created based on the median of each variable (1 if equal or greater than median, otherwise 0). The columns “day”, “month” and “year” are created from the “date” information.

3. Prepare for the word embedding. Convert the the reviews into tokens list or bigrams list by nltk.tokenize and nltk.collocations (BigramAssocMeasures and BigramCollocationFinder). For the tokens, STOPWORDS and numerical strings have been removed and then converted to stems. For the bigrams, STOPWORDS have been kept because they usually represent some meanings. Then, one-hot-vector for each review has been created based

on the vocabulary, which will be used for machine learning.

4. Build the prediction system by different algorithms include: Naive Bayes, Keras and LSTM. “is_effect”, “is_ease”, “is_satisfy” and “is_helpful” have been used as the label and train test split has been done before training the models.

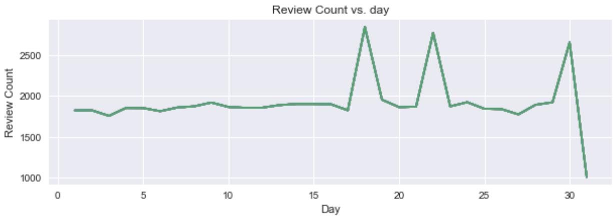
5. Text-mining to explore the most important phrases or terms that make the review impacting. Topic modeling has been implemented by Latent Dirichlet Allocation (by gensim library).

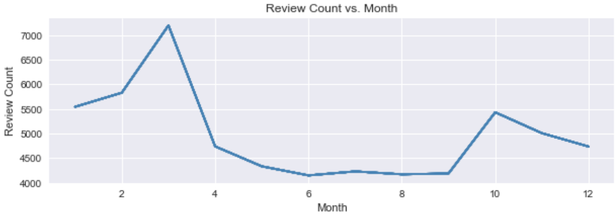
6. Another prediction system has been built based on the features created by topics. The algorithms for machine learning include: Naive Bayes, Random Forest and SVC.

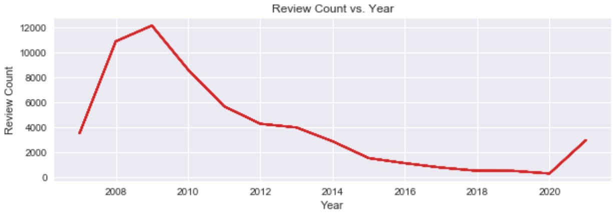
Evaluation and Results

Data Exploration

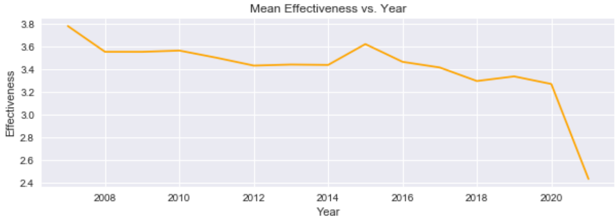
The plot of review counts by years, months and days have some interesting trends. The reviews posted in the second half month is more than that in the first half month. The highest review count occurs in March during the year which imply that people are easy to get sick in spring that improves the drug purchase. The review count kept decreasing from 2009 to 2019 and rebounds from 2020 when the COVID-19 pandemic occurred, which can be connected to the significant breakthroughs in medicine research of the decade, for example, genetic engineering obtaining major development, a vaccine and new treatments to fight Ebola, progress toward a vaccine for Dengue Fever, computer program predicts drug side effects and life extension breakthrough etc. These developments improved people’s health level and, in the meantime, increased people’s requirement on the qualities of medication.

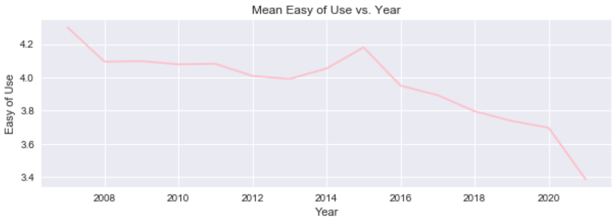


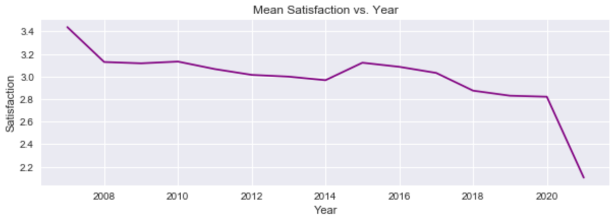




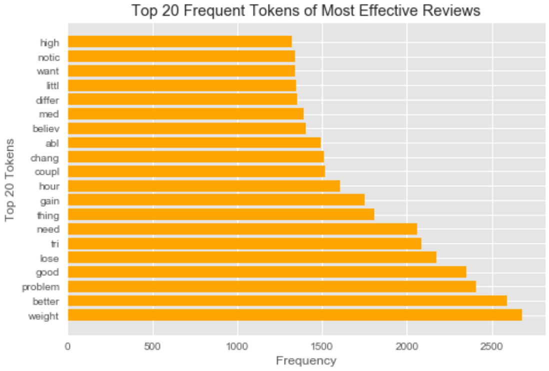
The plot of average “effectiveness”, “easy to use” and “satisfaction” indicate very consistent decreasing trends from 2007 to 2020, which means that people have gradually declining contentment on all kinds of medications and people’s demands become more and more strict.

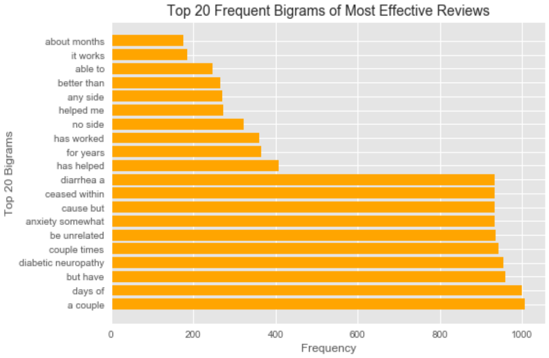


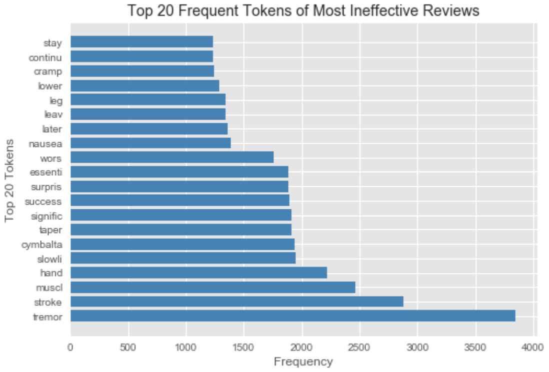


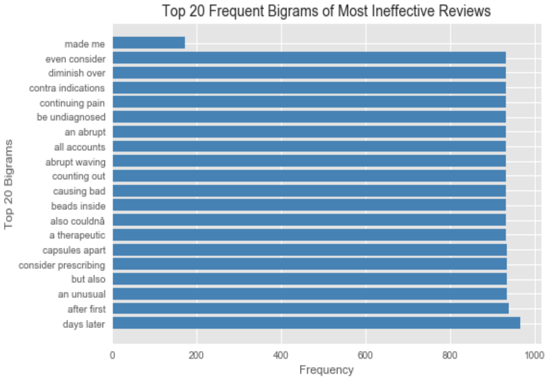


By analyzing the most frequent tokens and bigrams in the reviews with Effectiveness score 5 vs 1, we can also understand the patients demands in different aspects. For example, people emphasize time when thinking about effectiveness of a drug and usually have a bad impression on effectiveness when there is no pain relief, or the symptom is intensified. We can also see that bigrams can express more meaning than tokens.





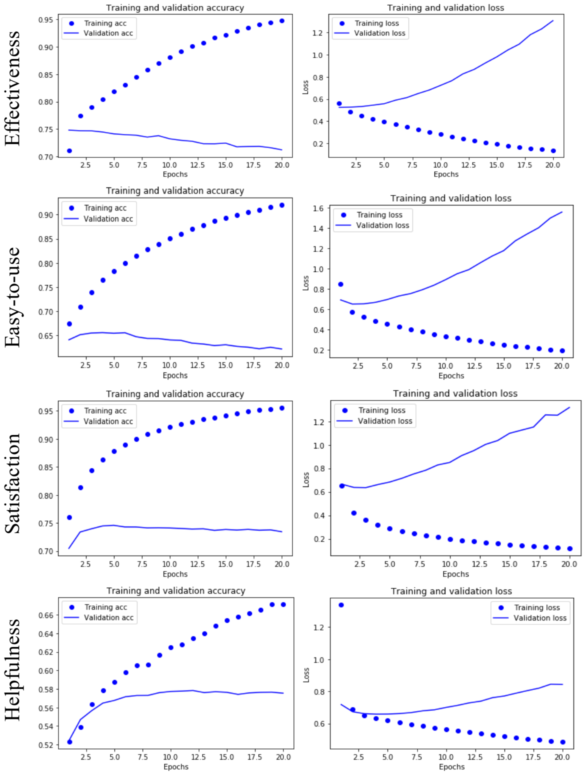




Prediction System by One-hot-vector

First, the tokenized data(unigram) has been used to build the training models. The vocabulary was created based on the most frequent 10,000 tokens without stopwords. Each review has been converted into one-hot-vector with 10,000 dimensions based on the vocabulary. The Naive Bayes model has been used as a baseline. For Naive Bayes model, the training matrix has been converted into a sparse matrix. In this study, there are four markers that can be used as the ground truth label for the classification/prediction: “is_effect”, “is_ease”, “is_satisfy” and “is_helpful”. F1-score has been used to evaluate the predictions.

Then a “keras” deep learning model has been tried to build the prediction system (two dense layers and one output layer ‘sigmoid’, the epochs were set to 20). The results show that, after the second epoch, the accuracy and loss for both train and test sets are deviated and there is overfitting between the train and test sets.



The accuracies and the f1-scores for all the labels are listed by the two tables below.

Compare the Naive Bayes model and “keras” deep learning model, we can see that the accuracy scores of the deep neural network are higher than Naive Bayes, while the F1-score of these two models are quite close, although the Naive Bayes model has a little bit higher F1-score. So, both of these models can make kind of reliable predictions.

The performance of LSTM deep learning neural network seems not so good as the regular "keras" model, it means that some of the hyperparameters are not tuned very well. So, this model needs to be improved.

Naive Bayes model

	Trainset Accuracy	Validation Accuracy	F1-score
"is_effect"	0.7122	0.6825	0.7643
"is_ease"	0.6705	0.6512	0.7315
"is_satisfy"	0.7344	0.7118	0.7317
"is_helpful"	0.5762	0.5535	0.6334

keras Deep Learning model (2 dense layers with one output layer)

	Trainset Accuracy	Validation Accuracy	F1-score
"is_effect"	0.9478	0.7122	0.7472
"is_ease"	0.9198	0.6222	0.6542
"is_satisfy"	0.9551	0.7343	0.7150
"is_helpful"	0.6711	0.5754	0.6252

LSTM Deep Learning model (3 dense layers with 3 dropout layer)

	Trainset Accuracy	Validation Accuracy	F1-score
"is_effect"	0.1834	0.1833	0.1879

Topic Modeling

The drug reviews of the entire dataset has been used for the topic modeling, and the method I used is LDA (by the gensim library). By training the review text data, 10 topics have been created.

Based on the keywords of each topic, I can extract the important medical elements or the aspects that patients care about from the topics, which are listed below:

Topic0: depression

Topic1: heart health and blood pressure

Topic2: depression and anxiety treatment

Topic3: dizzy and headache

Topic4: sleep and time

Topic5: weight loss

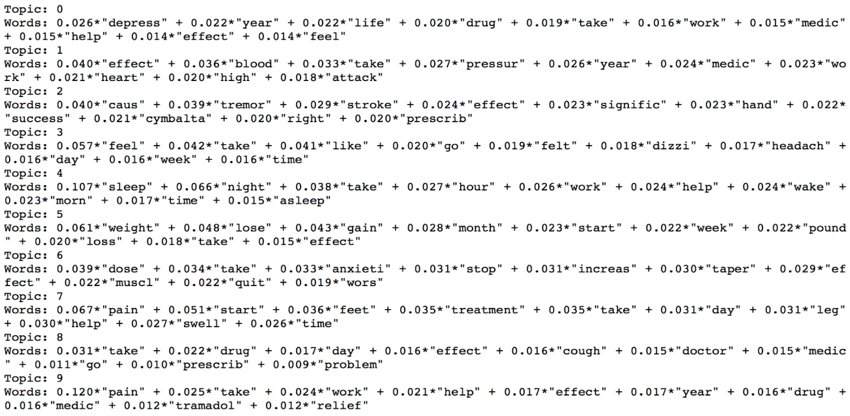
Topic6: anxiety treatment

Topic7: pain treatment

Topic8: cough and prescription

Topic9: pain relief

We can see that the main medical problems people encounter currently include depression and anxiety, heart disease, chronic pain, obesity and lack of sleep, etc. This kind of information provides important information to the pharmaceutical industry and medicine market.



```
Topic: 0
Words: 0.026*"depress" + 0.022*"year" + 0.022*"life" + 0.020*"drug" + 0.019*"take" + 0.016*"work" + 0.015*"medic"
+ 0.015*"help" + 0.014*"effect" + 0.014*"feel"
Topic: 1
Words: 0.040*"effect" + 0.036*"blood" + 0.033*"take" + 0.027*"pressur" + 0.026*"year" + 0.024*"medic" + 0.023*"wo
rk" + 0.021*"heart" + 0.020*"high" + 0.018*"attack"
Topic: 2
Words: 0.040*"caus" + 0.039*"tremor" + 0.029*"stroke" + 0.024*"effect" + 0.023*"signific" + 0.023*"hand" + 0.022*
"success" + 0.021*"cymbalta" + 0.020*"right" + 0.020*"prescrib"
Topic: 3
Words: 0.057*"feel" + 0.042*"take" + 0.041*"like" + 0.020*"go" + 0.019*"felt" + 0.018*"dizzi" + 0.017*"headach" +
0.016*"day" + 0.016*"week" + 0.016*"time"
Topic: 4
Words: 0.107*"sleep" + 0.066*"night" + 0.038*"take" + 0.027*"hour" + 0.026*"work" + 0.024*"help" + 0.024*"wake" +
0.023*"morn" + 0.017*"time" + 0.015*"asleep"
Topic: 5
Words: 0.061*"weight" + 0.048*"lose" + 0.043*"gain" + 0.028*"month" + 0.023*"start" + 0.022*"week" + 0.022*"pound
" + 0.020*"loss" + 0.018*"take" + 0.015*"effect"
Topic: 6
Words: 0.039*"dose" + 0.034*"take" + 0.033*"anxieti" + 0.031*"stop" + 0.031*"increas" + 0.030*"taper" + 0.029*"ef
fect" + 0.022*"muscl" + 0.022*"quit" + 0.019*"wors"
Topic: 7
Words: 0.067*"pain" + 0.051*"start" + 0.036*"feet" + 0.035*"treatment" + 0.035*"take" + 0.031*"day" + 0.031*"leg"
+ 0.030*"help" + 0.027*"swell" + 0.026*"time"
Topic: 8
Words: 0.031*"take" + 0.022*"drug" + 0.017*"day" + 0.016*"effect" + 0.016*"cough" + 0.015*"doctor" + 0.015*"medic
" + 0.011*"go" + 0.010*"prescrib" + 0.009*"problem"
Topic: 9
Words: 0.120*"pain" + 0.025*"take" + 0.024*"work" + 0.021*"help" + 0.017*"effect" + 0.017*"year" + 0.016*"drug" +
0.016*"medic" + 0.012*"tramadol" + 0.012*"relief"
```

Prediction System by LDA

Ten features of the drug reviews have been created by the 10 topics. Because there are 10 keywords in each topic with the corresponding weight, for each review text, the values of the features were created by summing the keywords weight in each review if the keywords occur in the review's token list. To compare with the one-hot-vector model, I implemented Naive Bayes on the topic dataset and calculated the f1-score. The result shows that the overall performance of topic features is not good as the one-hot-vector matrix, especially, the f1-score of "is_satisfy" and "is_helpful" is too low, which means that this method cannot be used to predict these classifications. But based on the f1-score of "is_effect" and "is_ease" which are close to the one-hot-vector matrix model, we can see that features extracted from top modeling can be used to make predictions in some conditions, which I have never seen in others reports.

Naive Bayes model on topic features

	Trainset Accuracy	Validation Accuracy	F1-score
"is_effect"	0.5947	0.5920	0.7384
"is_ease"	0.5632	0.5663	0.7173
"is_satisfy"	0.5327	0.5369	0.0241
"is_helpful"	0.5561	0.5580	0.0

Discussion

The goal of this study is to explore the patients' needs by analyzing the drug review text and build the predictions system to predict the drugs attribute such as "effectiveness", "easy-to-use", "satisfaction" and "helpfulness". The model accuracy and the f1-score shows that when using one-hot-vector and "keras" deep learning model, the performance of the prediction system is good for the prediction of "effectiveness", "easy-to-use", "satisfaction", but not so good for the prediction of "helpfulness" (that can be explain by the wider range of the helpfulness value). In addition, the topic modeling and data-mining analysis provide very interesting information about people's major health issues and people's demands on drugs, which would be very useful for new drug research and development and pharmaceutical marketing forecasts.

The baseline I used for the prediction system is Naive Bayes, which is the most common algorithm that is used as the baseline for text analysis. The evaluation metris shows that Naive Bayes can make good performance and would not take so much computing power. Then I used a deep learning model to improve the performance. The "keras" neural network has similar f1-score and higher accuracy than Naive Bayes, and the speed is fast. So, it can be used to do useful science. The LSTM is supposed to be the typical deep learning algorithm for text classification. But the performance is not good in this study and the training process is very time consuming. Thus, LSTM is not recommended for this case.

The f1-score in the range of 0.7 to 0.8 might not be considered as high, but because the data that are collected from 2007 till current year, and the reviews are posted by quite diverse people, so the f1-score in this range is good enough for this real case. And one-hot-vector and "karas" deep learning algorithms are most commonly used text classification method, it usually has good performance on text analysis.

Conclusion

The aim of this study is to build a prediction system to estimate the important attributes of drugs performance and uncover customers' demands by natural language processing techniques. We retrieved the patients' drug reviews and related information from WebMD (https://www.webmd.com) and used the "effectiveness", "ease-to-use", "satisfaction" and "helpfulness" ground truth labels to implement deep learning on the review text. The evaluation metrics indicate that this system is accurate to make the precision on "effectiveness", "ease-to-use" and "satisfaction". In the meantime, we used topic modeling to investigate the core elements from the review text and extract ten major topics, which reveals the most critical issues about people's health situation, for example body weight, sleep time, chronic pain and heart disease. This research will be useful for the pharmaceutical industry to understand customer’s needs and improve the product quality.

Other Things We Tried

In this study, one thing that took too much time is the web scraping. Because the web pages have been modified constantly, the web scraping was not very successful in the beginning, that made some results in the “project update” file not accurate. So, I had to redo the web scraping. Fortunately, the updated data have been collected correctly and successfully by the recent trial, and the newest data have been collected.

Additionally, I tried to implement classification based on the features created from topic modeling and I have tried lots of different machine learning algorithms. Although the f1-scores show that the “satisfaction” and “helpfulness” cannot be predicted well by this approach, the result for “effectiveness” and “ease-to-use” is not bad. Because I have never seen the same approach used in other reports, I think this method is a new approach that can be used to build classification models. Although I don’t have very surprising results in this case, I’m sure it is worth trying in some other study.

What You Would Have Done Differently or Next

Beside the attributes analyzed in this study, eg “effectiveness” and “ease-to-use”, there is an important element that has not been studied in this research, “condition”. To better understand people’s health needs, predicting the conditions by the patient’s posts is a good approach. Because the “condition” variable includes very diverse values, such as “depression”, “hypertension” or “Schizophrenia”, that make the classification very complicated. We need more instances and more capable algorithms to fine-tune the model.

Another aspect that I think I can make improvements is trigram. By doing text-mining on tokens and bigrams, I found that bigram can express more meanings than tokens. Thus, trigrams or longer phrases might be more informative and representative for people's opinions. So, I plan to work on trigrams for the next step.

Acknowledgments

The first person I want to thank is our professor Dr. David Jurgens. He gave us the best lectures about natural language processing and inspired me to do more research on this field in my career. This course prepared me better to look for a Data Science job in the future.

And I also want to thank our GSI, Zhuofeng Wu. He gave me lots of help on my homework and project. He is always very patient on all kinds of questions and he provided me with very helpful suggestions on my problems.

References

[1] Sentiment analysis using product review data. Xing Fang and Justin Zhan, Journal of Big Data, 2015, 2:5 DOI 10.1186/s40537-015-0015-2.

[2] Towards Enhanced Opinion Classification using NLP Techniques. Akshat Bakliwal, Piyush Arora, Ankit Patil, Vasudeva Varma, Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP 2011, pages 101–107.

[3] Classification of Book Reviews Based on Sentiment Analysis: A SURVEY. A. Mounika,Dr. S. Saraswathi, International Journal of Research and Analytical Reviews, 2019 IJRAR June 2019, Volume 6, Issue 2.

[4] Towards Enhanced Opinion Classification using NLP Techniques. Akshat Bakliwal, Piyush Arora, Ankit Patil, Vasudeva Varma, Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP 2011, pages 101–107.

[5] Text-mining of PubMed abstracts by natural language processing to create a public knowledge base on molecular mechanisms of bacterial enteropathogens. Sam Zaremba, Mila Ramos-Santacruz, Thomas Hampton, Panna Shetty, Joel Fedorko, Jon Whitmore, John M Greene, Nicole T Perna, Jeremy D Glasner, Guy Plunkett III, Matthew Shaker David Pot, BMC Bioinformat- ics, 2009, volume 10, Article number: 177.

[6] Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. Nikolaos Korfiatisa, Panagiotis Stamolamprosb, Panos Kourouthanassisc, Vasileios Sagiadinos, Expert Systems with Applications, Volume 116, February 2019, Pages 472-486

[7] Automated topic modeling of tourist reviews: Does the Anna Karenina principle apply? Andrei P. Kirilenkoa, Svetlana O. Stepchenkovaa, Xiangyi Dai, Tourism Management, Volume 83, April 2021, 104241

[8] Sentiment Analysis of Product Reviews: A Review. Shivaprasad T. K., Jyothi Shett, International Conference on Inventive Communication and Computational Technologies, (ICICCT 2017)

[9] Customer Perception Analysis Using Deep Learning and NLP. Sridhar Ramaswamy, Natalie DeClerck,Procedia Computer Science 140 (2018) 170–178.

[10] Very quaffable and great fun: Applying NLP to wine reviews. Iris Hendrickx, Els Lefever, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, August 7-12, 2016, pages 306–312