Using Machine Learning and Sentiment Analysis to Assess Pharmaceutical Product Sentiment

Frank Cuoco Jr.
University of California, Berkeley
Computer Science, Class of 2022
fcuoco@berkeley.edu

Abstract

Objective:
Healthcare-related social media sites such as Drugs.com and WebMD are rapidly becoming a source of real-time information on patients' experiences with pharmaceutical products. With millions of these user reviews online, it would be valuable to analyze the content of these reviews to understand patients' perception and sentiment (positive, negative, or neutral) regarding these products. Much research has been done on using machine learning to do this on simple and succinct types of pharmaceutical reviews from sources such as Twitter and Facebook, but more complex, multi-sentence, and "story-like" reviews have presented challenges to researchers.

The objective of this project is to utilize machine learning techniques to build a generalized model capable of accurately identifying the sentiment of pharmaceutical drug reviews of any length and complexity, in order to more fully understand the overall market perception of a drug.

Methods:
To initiate the project, a database was populated by scraping over 6,200 pharmaceutical drug reviews from 55 of the top selling office-based medications across 13 pharmaceutical classes using Drugs.com, a health related social media website. The data was then preprocessed, balanced, stratified, and split into training and test datasets in preparation for classification using several machine learning classifiers such as Random Forest (RF), Multinomial Naïve Bayes (MNB), and Linear Support Vector Machines (SVM).

Results:
Hybrid Linear SVM was the most accurate at predicting sentiment of reviews across all drug therapeutic areas, with a training accuracy score of 85.6% and a testing accuracy score of 74.1%.

Conclusions:
The results indicate that a generalized model could be developed using machine learning methods to identify patient sentiment for pharmaceutical products over a wide variety of therapeutic classes. Future areas of research can include methods to increase model accuracy even further, as well using the model to evaluate the correlation between sentiment and overall pharmaceutical product success.

1. Introduction

The proliferation of social media sites like Facebook and Twitter and the overall growth of online forums has revolutionized the information age, allowing social scientists a plethora of user reviews and comments to analyze. Being able to actually utilize this rapidly growing amount of big data, however, comes with its associated challenges. One challenge with big data is that the sheer size and continual growth of it makes manual analysis by humans infeasible.

Another challenge is classifying this data. In fact, studies by Wilson, et al. (2005) and Saif, et al. (2013) have shown that that even human readers could only come to agreement 65% to 80% of the time when trying to manually classify comments/reviews. Reports from software developers like Lexalytics (2019), who develop and market automated sentiment analysis systems, have also echoed similar results in their testing of human accuracy vs. automated systems.

As a result of these shortcomings, sentiment analysis has gotten increased attention as a possible solution for this problem. Also known as opinion mining, sentiment analysis is a process that utilizes natural language processing and other methods to allow computers to automatically identify and classify the intended sentiment of human generated text (whether it is positive, negative, or neutral). Sentiment analysis has seen many advances recently in its ability to successfully classify such text with machine learning techniques, and with its current rate of growth, it is clear that this process has the ability to approach or exceed human accuracy levels Kharde, et al. (2016).

Early sentiment analysis research focused in on consumer reviews in popular consumer package goods markets (Pang, et al. 2008 and Liu, et al. 2012), and has only in the past five years begun to be used in the processing of medical and pharmaceutical information from patients. While initial studies involved the detection of adverse drug events (ADEs) in pharmaceutical markets, more recent studies have focused on experimenting with sentiment analysis on patient reviews of drug experiences coming from general social media websites such as Twitter and Facebook.

While much work has been done in this area, this work tended to be focused on shorter and less complex review texts. Later work focused on the same type of patient reviews, but instead coming from more forum-based, health related sites such as Drugs.com and WebMD. The main difference between this work and the previous work was that reviews coming from these health related websites were significantly more detailed, consisting of nearly three to four times the number of sentences than those from traditional social media sources, and were formatted in a verbose, story-like way. This made analyzing reviews from these sources a much more difficult task to undertake.

2. Literature Review

Various approaches over the past 15 years have been developed for sentiment analysis. One of the most basic approaches, “Bag of Words”, analyzes groups of comments / reviews and extracts key words to train the model. The assumption of this approach is that if the vocabulary and the corresponding polarity of a sample group of comments/reviews are captured, then one will be able to use this information to identify the polarity of any new comments or reviews that were not originally used to build the model. Besides this, machine learning techniques such as Naïve Bayes, Support Vector Machines (SVM), and Random Forest classifiers are also commonly used. Lastly, hybrid approaches that leverage the above methods or other multi-stage or decision rule-based components are potential approaches that have been investigated Ghiassi, et al. (2013).

In the pharmaceutical area, several studies of sentiment analysis have attempted to evaluate complex patient drug reviews with various levels of success, and initially focused on identifying adverse drug event prevalence mentioned by patients. Adverse Drug Events (ADEs) are unexpected significant side effects and/or safety issues that emerge after a product is approved by the FDA and is used by the general public. In many of these situations, the ADEs were so severe that the product was eventually withdrawn from the market.

Work conducted by Chee et al. (2011) used machine learning classifiers to identify products with patient messages that were consistent with messages from drugs that were eventually withdrawn from the market due to adverse drug events. Fang et al. (2013) took a different approach, which involved building a custom lexicon of the current drug’s existing side effects, then comparing this with pharmaceutical reviews to categorize and identify unrecognized drug side effects (likely ADEs). Lastly, Sarker et al. (2015) proposed a method of using machine learning to automatically detect mentions of ADEs. It was based on corpus training of 74 drugs in nine therapeutic categories, resulting in a high number of features used to identify products with significant ADEs. Because of the breadth of products in this model’s corpus, it showed potential for being able to be used to evaluate ADEs for any pharmaceutical product.

In addition to identifying pharmaceutical product ADEs, there has also been a need to identify patient sentiment regarding their pharmaceutical drug experiences from online reviews. This would allow pharmaceutical manufacturers to obtain valuable customer feedback, understand any misconceptions or issues with the use of their products, and to be able to make changes and respond in a timely and effective manner to ensure the success of these products.

Na et al. (2015) utilized a clause-level linguistic based sentiment analysis algorithm based on data from WebMD in five therapeutic categories containing over 4,200 sentence clauses from drug reviews to determine patient sentiment. Data was also analyzed using the SVM machine learning approach. Results showed the linguistic approach slightly outperformed the SVM approach in terms of accuracy, with 69% versus 66%, respectively.

Gopalakrishnan et al. (2017) used a Radical Basis Function (RBF) neural network to analyze drug reviews for two pharmaceutical products to determine sentiment. RBF methods were also compared to SVM and probabilistic neural network (PNN) methods. RBF was found to be superior in this analysis with an 89.1% accuracy vs. 86.5% for PNN and 80.1% for SVM. However, since this study was based on only two specific pharmaceutical products, its broader use and applicability was limited.

Mahoob et al. (2018) used a lexicon approach to evaluate patient reviews of pharmaceutical products from two patient drug evaluation websites, livewell.pk and kaymu.pk. This lexicon approach utilized SentiStrength, which was an application developed by researchers Thelwall et al. (2010) and was based on the analysis of 2,600 comments from the social media site MySpace, a precursor to sites like Facebook. Using this, 25 pharmaceutical products in nine therapeutic areas were evaluated with an average precision score of 54%, which is a measure of accuracy of correctly classifying reviews. The only drug categories that scored above the model average were Eye/Skin, Sexual Wellness, and Allergy/Sinus. The remaining categories of Children's Healthcare, Dehydration, Digestive Health, Pain and Fever, Cough/Cold, and Women's Health all scored well below the model average of 54%, effectively making these markets unreliable in terms of model predictability.

In summary, studies specifically focused on determining the patient sentiment from drug reviews have been conducted by several researchers, with varying levels of success and applicability. Moreover, most results were based on a small sample size of pharmaceutical reviews among a limited number of therapeutic classes. These studies serve as a good starting point, but make apparent the need for a more comprehensive study with larger sample sizes and a resulting model with more general applicability.

Thus, the objective of this research was to build a generalized model with sufficient breadth and depth among key pharmaceutical therapeutic areas that could be used to evaluate the sentiment of any pharmaceutical product regardless of therapeutic category.

3. Methods

3.1 Datasets

With the goal of building an accurate, generalized sentiment model for drug reviews in mind, a balanced dataset with a diverse number of medications, specifically ones that are representative of the current pharmaceutical market, is crucial. In order to achieve this, a dataset with over 6,200 pharmaceutical drug reviews from 55 of the top selling office-based medications across 13 pharmaceutical classes was compiled.

Drug classes included were: Anticoagulants, Asthma, Cardiovascular, Cholesterol, Central Nervous System (CNS) Disorders, Diabetes, Gastrointestinal Disorders, Hepatitis-C, HIV, Pain, Respiratory Disease, Sexual/Urological Health, and Smoking Cessation (Figure 1). CNS Disorders made up the largest segment of the data (22%), due to the fact that it is an aggregation of many sub-categories like depression, anxiety, epilepsy, and Parkinson's Disease.

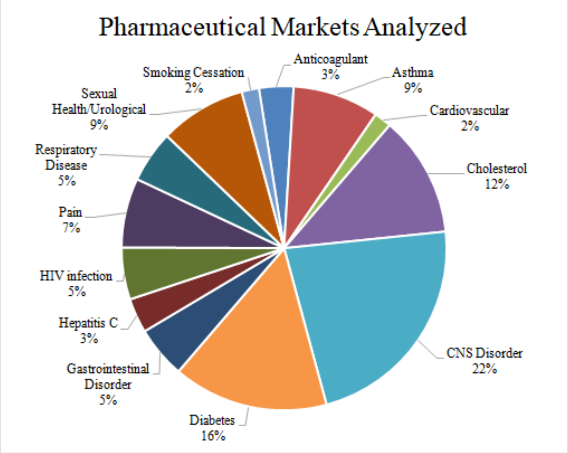


Figure 1. Drug Class Distribution in Sample

In the pharmaceutical domain, there are numerous social websites where people can discuss and review various drugs they have taken, and it was ultimately decided to use

Drugs.com for populating the dataset because their reviews were more comprehensive and complex than competing websites. The reviews also had ratings attached to them that used a 1 to 10 scale, with 1 representing a very negative experience, and 10 a very positive experience. Reviewers were instructed to give their rating based on how effective they found the medicine, taking into consideration efficacy, side effects, and ease of use. These ratings would allow the training data to be more quickly labeled with polarity values needed for this analysis.

3.2 Data Collection

Python methods were used to scrape Drugs.com data and export key review and ratings information to an Excel database. Although information such as the date of the review, the username of the reviewer, and the period of time the reviewer was on the medication was collected, only information necessary for sentiment analysis was analyzed, that being the text of the review and the numerical rating. Again, an important distinction between this study and others before it is that unlike tweets, these reviews contain multiple, complex sentences. Many times, a review will contain both negative and positive sentences, and it is up to the model to distinguish what the overall sentiment of the review actually is, which can be challenging. This will be looked at in greater depth in the results section. An average excerpt of scraped reviews is shown below in Table 1.

Table 1.
MASTER DATA: Reviews

Drug	Polarity	Review
Lipitor	Negative	"My cardiologist took me off Lipitor 2 weeks ago-my cholesterol is great after being on a low carb diet. Previously my body was covered in bruises at all times which I had attributed to other medicines (Plavix and aspirin) I Just realized I am bruise free for the first time in over a year. Conclusion: it was the Lipitor causing the problem."
Lunesta	Neutral	"Horrible aftertaste that lasts at least 8 hours. The taste comes about 5 mins after I take it and that's when I know it's working. Otherwise I can't tell like I could taking another popular sleeping medicine. It's increased my depression and anxiety. I'm getting put back on my previous medication in three days due to the horrible after taste. It's chemical tasting. Nasty."
Xolair	Positive	"I've been on this medication since August 2014. It really has improved my quality of life. Before it, I was in and out of the doctors or ER every week. But now I rarely need to go because of this medication. I have tried so many different medicines and this is the only one that worked for me. It's scary when you see your lung function, decreasing over time But Xolair has given me hope. I get the shot every 2 weeks. Sometimes the side effects suck, muscle weakness and extreme fatigue but it's better then not breathing"

Note. Reviews were categorized based on a 1 to 10 rating scale, where 1-4 is considered "negative", 5-6 is considered "neutral", and 7-10 is considered "positive".

Pharmaceutical reviews have one flaw when being used for sentiment analysis. Due to the nature of the people who make reviews on health-related social media sites like Drugs.com, there tends to be a larger volume of positive reviews vs. negative and neutral reviews. In this case, neutral reviews initially were nearly half the amount of negative reviews. As a result of this imbalance in reviews, oversampling was performed to normalize the neutral reviews, resulting in a final sample size of 6,262 reviews, as shown in Table 2.

Table 2.
DRUG SENTIMENT: Reviews by Sentiment

Polarity	Number of Reviews
Positive	3,593
Negative	1,307
Neutral	1,362
Total	**6,262**

3.3 Preprocessing of Data

In order to prepare the data for the analysis, patient reviews needed to first be preprocessed. After parsing out the reviews word by word, the first step was to strip away stop words from each review. Stop words are words commonly used that don't offer additional help in sentiment identification such as "and", "the", "at", "not", "any", etc. However, since certain stop words like "not" and "any" are in many scenarios important in identifying sentiment of drug reviews (as is the phrase "there were not any side effects"), important stop words like these were not removed from the reviews.

All words were then made lowercase and all numbers were removed for consistency purposes. Finally, the reviews were lemmatized, which means all tenses of a word are condensed into one term. This prevented words from being overcounted if they showed up in several different tenses throughout a review (e.g. "works", "worked", "working", converted to → "work"). These processes were then compiled into a custom normalizer function in Python, which had the ability to take in raw reviews and output normalized reviews. Table 3 shows an example of this process on the first few sentences from each of the reviews.

Table 3.
PREPROCESSING: Normalizing Reviews

Input	Output
"My cardiologist took me off Lipitor 2 weeks ago."	"cardiologist took off lipitor week ago"
"Horrible aftertaste that lasts at least 8 hours."	"horrible aftertaste last at least hour"
"It really has improved my quality of life."	"really improved quality life"

With the reviews normalized, features can now begin to be extracted from each review. A feature is an aspect of a processed piece of data that can be used to compute the classification of further data, and is a required input in order to use machine learning. In the case of sentiment analysis, the features are words or groups of words extracted from each review that will be used to compute the polarity of future reviews. These groups of words are called n-grams, and specifically unigrams and bigrams (one and two word pairs) were evaluated as potential feature vectors for the analysis. A hybrid approach (using both unigrams and bigrams) was used because in health-related reviews, while words like "horrible" and "terrible" are strong indicators of sentiment by themselves, sometimes the context of a word matters and can entirely change its perceived sentiment, requiring the word before or after it to be included. For example, the word "pain" sounds negative when by itself (unigram), but if one looks at the word preceding it, the phrase could actually be "no pain" (bigram), which completely changes its polarity from negative to positive. Furthermore, the word "works" (unigram) could be positive if it was in the phrase "it works" (bigram), but could be negative if it was part of the phrase "works poorly" (bigram). Figure 2 shows the effect of using a hybrid n-gram approach on the machine learning method of Linear SVM, and clearly shows that a hybrid approach performs better than the other n-grams approaches. This increase in performance of the hybrid n-gram approach was also observed across all classification methods.

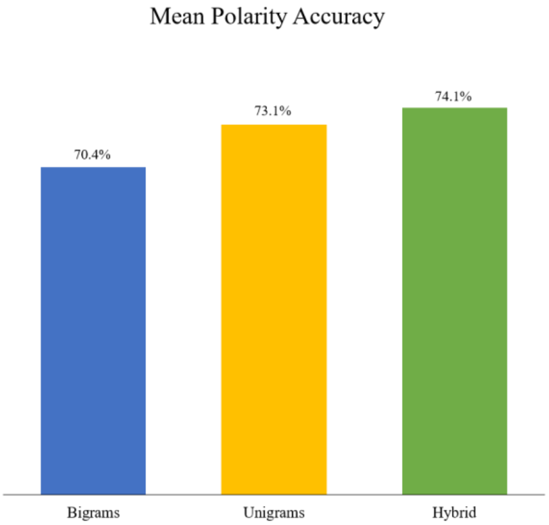


Figure 2. Linear SVM Model Accuracy Based on N-Gram Approach

An example of this hybrid process of extracting unigrams and bigrams is shown on the reviews in Table 4.

Table 4.
FEATURE EXTRACTING: Extracting Features from Normalized Reviews

Normalized Review	Features (Unigrams & Bigrams)
"cardiologist took off lipitor week ago"	"cardiologist", "took", "off", "lipitor", "week", "ago", "cardiologist took", "took off", "off lipitor", "lipitor week", "week ago"
"horrible aftertaste last at least hour"	"horrible", "aftertaste", "last", "at", "least", "hour", "horrible aftertaste", "aftertaste last", "last at", "at least", "least hour"
"really improved quality life"	"really", "improved", "quality", "life", "really improved", "improved quality", "quality life"

3.4 Splitting the Data into Test and Training Datasets

Before the model can be trained, the data must be split into training (80%) and test (20%) datasets. To do this, stratified random sampling was performed to ensure that both datasets were balanced, with each drug therapeutic area having proportional representation within both datasets. These two datasets were then used to train and test various machine learning classifiers.

3.5 Methods for Sentiment Analysis Using Classifiers

Based on previous research on sentiment analysis, Mullen et al. (2004), Jadva et al. (2016), and Kharde et al. (2016), three key classification methods were identified as being most useful in this analysis: Random Forest, Multinomial Naïve Bayes, and Support Vector Machines.

Random Forest classification (RF) utilizes the tree data structure to ultimately classify data. The classifier takes an ensemble (or “forest”) of random and uncorrelated binary decision trees that are created through “bootstrap aggregation” and averages their decisions to come up with a “majority-vote”. The less correlation between each tree, the more accurate the model. One major issue with RF is that the deeper each tree is grown (i.e. the more branches each decision tree has), the more likely the model will overfit its training data, causing major issues with its predictive power on new data.

Multinomial Naïve Bayes (MNB) is a model based on probability that uses Bayes mathematical theorem and naïve assumptions to classify data. An advantage of MNB is that it only requires a small amount of training data and has a short training time, allowing a minimal amount of memory to be used compared to other methods like RF and SVM. However, because of its simplistic technique, it is usually the least accurate of the three. MNB was chosen over regular Naïve Bayes as it performs better with text-based data due to its ability to track the frequency of a word instead of just its presence or absence.

Support Vector Machine (SVM) is a plane-based machine learning method that seeks to find a hyperplane where the equal distance between a set number of classes of data points is maximized. In the case of this study, there are three clusters of data points: “positive”, “negative”, and “neutral”. The more distance between the clusters, the more accurate future points can be modeled. Previous research, Mullen et al. (2004) and Jadva et al. (2016), suggests that SVM methods, especially hybrid SVM, have been found to be one of the most accurate classifiers in sentiment analysis.

4. Results

Table 5 shows the accuracy of both the training and test datasets for the three machine learning methods tested. The training accuracy metric in the first column indicates how well the model fit the training dataset, and Linear SVM performed the best with 85.6% accuracy while Naïve Bayes (71.0%) and Random Forest (59.3%) performed significantly worse. The more important metric for evaluating the effectiveness of each classifier is the test accuracy metric.

Since none of the data in the test dataset was used in the training of the model, the test accuracy metric truly illustrates the classifiers’ true predictive power. Results here indicate that Random Forest and Multinomial Naïve Bayes performed the poorest, with accuracies of 58.2% and 62.7%, respectively, while Linear SVM was clearly the most effective in predicting sentiment of the three, scoring well at 74.1% accuracy. The difference between the accuracy of the training and test data for Linear SVM is an indicator that the model is doing a relatively good job of correctly classifying new data. As a result, the Linear SVM classifier was chosen for this analysis.

Table 5.
MODEL ACCURACY: By ML Classifier

Machine Learning Classifier	Training Accuracy	Test Accuracy
Random Forest Classification	59.3%	58.2%
Multinomial Naïve Bayes	71.0%	62.7%
Linear SVM with TF-IDF Vectorization[a]	85.6%	**74.1%**

Note. Sample of 6,262 reviews

[a] TF-IDF Vectorization is a way of vectorizing reviews that takes into consideration how frequently a word appears in a review.

As shown in Table 6, besides overall accuracy, the SVM model was also evaluated by using the precision metric, split by sentiment class (positive, negative, and neutral). Precision is the number of correctly classified reviews of a class divided by all the reviews of that same class, or essentially it is accuracy split by class.* Positive reviews were predicted best with a precision of 89%, with negative reviews less at 62%, and neutral reviews slightly lower at 59%.

Table 6.
MODEL PRECISION: By Review Polarity

Polarity	Precision
Positive	89%
Negative	62%
Neutral	59%

*see Appendix for detailed formula on precision calculation

The reason for the disparity in precision by class may be due to several factors. For instance, since the dataset had twice as many positive reviews than negative and neutral reviews, the model had significantly more positive reviews available for training, possibly resulting in better positive polarity predictions.

Another major factor is that positive reviews tend to be more concise and straightforward, while negative and neutral reviews tend to be more complex. The reason for this is that if the drug is satisfactory and works, the patient is able to summarize their positive feelings in a more direct way. It is only once there is a list of issues to complain about (side effects, comparing the drug to others, etc.) that the patient begins to complicate the review by adding additional sentences about their many qualms with the medication over the period of time they attempted to use it. As a result, the model will do much better predicting the positive reviews than the negative/neutral reviews. Table 7 shows an example of this as Chantix, a smoking cessation agent, receives overwhelmingly positive reviews on Drugs.com. The drug simply works, and the patient has nothing more to say about the it, causing the review to be simple and concise, and thus easily identifiable.

Table 7.
MASTER DATA: Example of Simple Review

Drug	Polarity	Review
Chantix	Positive	"I started Chantix four weeks ago and I have completely stopped smoking after 15 years. This is a great product! I slowly dwindled down on the amount I smoked per day until my body just didn't want a cigarette anymore. The only side effect I had was headaches in the beginning. I'm on an antidepressant as well. I feel great and am so thankful for this medication. The smell of cigarettes are horrible now! Yay"

This is in comparison to CNS drugs that are for more complicated issues and are more likely to vary greatly in effectiveness from person to person. For example, an anxiety medication may reduce a user's anxiety, but its side effects could then cause insomnia and heavy withdrawal effects if the user misses a day of medication. It is easy to see how that could make the reviews somewhat longer and more difficult to analyze. Instead of getting straight to the point on if they like or dislike the medication, many users will instead spend several sentences talking about their "rollercoaster journey" with their disorder, and say nothing about the actual medication they are reviewing or anything that could be useful in identifying sentiment. To add to this, once the user begins to talk about the medication, they many times will compare it to the myriad of other drugs that they have taken previously, potentially confusing the model.

For instance, in the review in Table 8, the user starts by saying drugs like Hydrocodone, Vicodin, and Valium work “good” for them, but the current drug they are reviewing (Vesicare) is “bad”, which seems contradictory when using ML methods to determine the overall sentiment of a review. Many of the drugs that fall into this category are negative and neutral reviews, which is most likely why their overall precision scores are lower, as the model is having trouble at times differentiating one for the other. This is reflected in their similar precision scores.

Table 8.
MASTER DATA: Example of Lengthy Review

Drug	Polarity	Review
Vesicare	Negative	“I took it for a few months and felt like a zombie. I've been dealing with stress and for some lame ass reasons, a health care shrink diagnosed me as bi-polar. All who know me would say that's a load of crock. Hydrocodone/Vicodin/Valium works well with me, but no Dr.s want to perscribe it to me. They are afraid I will abuse it. I've never abused any medication, nicotine, alcholol etc. in all my 56 years on this planet. I just know what I need/want/enjoy for the time I need it. It won't fix my problems, but It certainly helps me" fix my problems. Lately, I've been getting Vicodin off the street. I take one a day. (7.5 mg) ...occasionally two. I am more positive and one by one am getting those pesky stressful situations out of my life.”

Another noteworthy split to look at is how the model performed by therapeutic area. As shown in Figure 3, drugs in the Smoking Cessation, Respiratory Disease, Anticoagulant, and Sexual Health/Urological areas performed better than the model average and were the easiest for the model to predict, while drugs associated with Cholesterol, Hepatitis C, and Diabetes were the most difficult for the model to classify.

The reason for this is most likely due to the multifaceted nature of certain diseases and their respective complex reviews. For instance, a patient trying to quit smoking likely is usually focusing on that single issue, whereas diabetes patients tend to have several other diseases in addition to diabetes (High Cholesterol, Hypertension, etc.), meaning that their reviews for diabetes medications often contain a lot of extra “noise” about the other medications they are currently taking and how they compare. This muddles the sentiment of the review and makes them more convoluted and harder to classify.

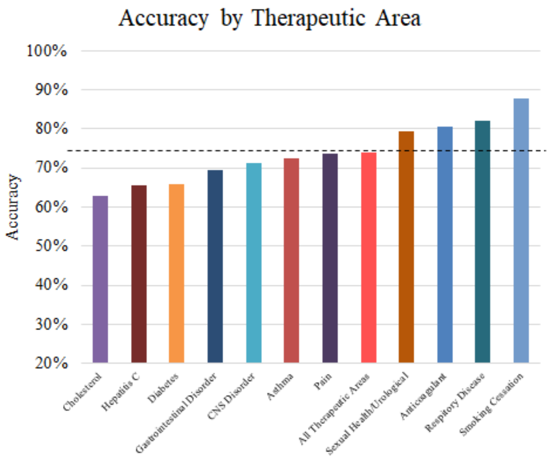


Figure 3. Model Accuracy by Therapeutic Area

5. Discussion

The findings indicate that machine learning can be used in sentiment analysis to effectively and efficiently predict patients’ feelings regarding their usage of various pharmaceutical products across a wide range of therapeutic areas. The SVM model used dhad a predictive overall accuracy rating of nearly 75% and was very precise in predicting positive reviews at 89%.

However, prediction of negative and neutral reviews were below the model average due to the complexity of these reviews, which made classifying them a challenge.

A key beneficial use of this model is its ability to quickly identify the sentiment of an entire forum of patient reviews, meaning it could be invaluable in helping pharmaceutical marketers understand patient satisfaction with their products in real time. Since the dawn of big data, there is simply too much information for companies to extract sentiment from manually, and a model like this would allow them to do that almost instantly. In addition, this model could be adapted to be able to track the perception of a certain drug so companies could evaluate the perception of their product over time versus other competing drugs, or evaluate negative/positive events or publicity. This analysis would allow management to analyze changes in market trends, respond quickly, and use this information to formulate strategic decisions.

Several further steps could be taken to improve this model. First, the pre-processing of reviews can be streamlined and made more sophisticated to include things like identifying key words that will always predict a certain polarity, or by giving confidence weights to each sentence in a given review and then averaging them out to make a final decision. Additionally, much work needs to be done to better predict lengthy negative and neutral reviews.

To do this, developing and implementing a multi-step procedure could eliminate the extraneous “noise” from these reviews and could help to further clarify the overall direction of the review, increasing the accuracy of the model. This could primarily be done by evaluating sentences individually and filtering the data of extraneous products and subjects that do not directly relate to the product being reviewed.

In order to truly reach the final goal of creating a generalized model, the model should be further improved and adjusted by continuing to train it with more data from other health related medical review websites (e.g. WebMD, RxList, etc.), medical forums, or even large databases like Twitter and Facebook. This would allow the model to eventually be used to assess patient perception on pharmaceutical products, regardless of source.

Lastly, this project could be a catalyst for new research projects. Future research could also involve using the approach taken in this project to create new models that can be used to predict sentiment of reviews from any field or industry outside of pharmaceuticals, such as film, technology, consumer products, etc. Another interesting area of future research could involve investigating the possible connection between the polarity trends of a new product and its impact on the product’s overall future success in the market.

6. References

[1] Chee, B. W., Berlin, R., & Schatz, B. (2011). Predicting adverse drug events from personal health messages. *AMIA ... Annual Symposium proceedings. AMIA Symposium, 2011,* 217–226.

[2] Fang, H., Stanhope, S. J., & Wu, H. (2013). Exploiting Online Discussions to Discover Unrecognized Drug Side Effects. *Methods of Information in Medicine, 52*(02), 152-159. doi:10.3414/me12-02-0004

[3] Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications, 40*(16), 6266-6282. doi:10.1016/j.eswa.2013.05.057

[4] Gopalakrishnan, V., & Ramaswamy, C. (2017). Patient opinion mining to analyze drugs satisfaction using supervised learning. *Journal of Applied Research and Technology, 15*(4), 311-319. doi:10.1016/j.jart.2017.02.005

[5] Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 04*. doi:10.1145/1014052.1014073

[6] Jadav, B., & Vaghela, V., (2016). Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis. International Journal of Computer Applications, 146(13), 26-30. doi:10.5120/ijca2016910921

[7] Kharde, V., & Sonawane, S. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. *International Journal of Computer Applications, 139*(11), 5-15. doi:10.5120/ijca2016908625

[8] Lexalytics, Inc. (2019, July 12). Sentiment Accuracy: Explaining the Baseline and How to Test It. Retrieved from https://www.lexalytics.com/lexablog/sentiment-accuracy-quick-overview

[9] Liu, B. (2012). *Sentiment analysis and opinion mining*. San Rafael, CA: Morgan and Claypool.

[10] Mahboob K, Ali F (2018). Sentiment Analysis of Pharmaceutical Products Evaluation Based on Customer Review Mining. *Journal of Computer Science & Systems Biology, 11*(3). doi:10.4172/jcsb.1000271

[11] Mullen, T., Collier, N., (2004). Sentiment analysis using support vector machines with diverse information sources. *Proceedings of Conference on Empirical Methods in Natural Language Processing, 2004*.

[12] Na, J., & Kyaing, W. Y. (2015). Sentiment Analysis of User-Generated Content on Drug Review Websites. *Journal of Information Science Theory and Practice, 3*(1), 6-23. doi:10.1633/jistap.2015.3.1.1

[13] Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval, 2*(1–2), 1-135. doi:10.1561/1500000011

[14] Saif, H., Fernandez, M., He, Y. & Alani, H. (2013). Evaluation Datasets for Twitter Sentiment Analysis. A survey and a new dataset, the STS-Gold. *CEUR Workshop Proceedings. 1096*.

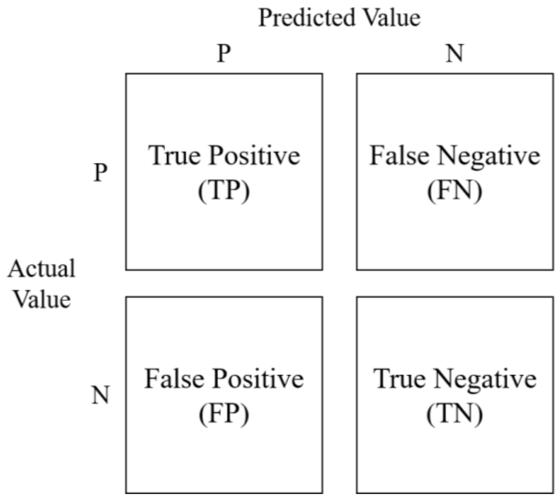
[15] Sarker, A., & Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics, 53*, 196-207. doi:10.1016/j.jbi.2014.11.002

[16] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544-2558. doi:10.1002/asi.21416

[17] Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT 05*. doi:10.3115/1220575.1220619

7. Appendix

Accuracy is one of four useful metrics that can be used to evaluate the effectiveness of this model, and they all stem from the confusion matrix. The confusion matrix (Figure 4) is extremely useful in measuring the performance of a classification model, and does this by relating True Positives / Negatives (TP & TN) and False Positives / Negatives (FP & FN) together.



		Predicted Value	
		P	N
Actual Value	P	True Positive (TP)	False Negative (FN)
	N	False Positive (FP)	True Negative (TN)

Figure 4. Confusion Matrix

Accuracy is the total number of correctly predicted reviews (TP + TN) divided by the total number of reviews (TP + TN + FP + FN). Precision is the total number of correctly identified positive reviews (TP) divided by the total number of predicted positive reviews (TP + FP). Next, recall is the total number of correctly identified positive reviews (TP) divided by the total number of actual positive reviews (TP + FN). Finally, F1 score is a weighted average of precision and recall. The formulas for all four metrics are shown in Figure 5.

$$\text{Accuracy} = \frac{\text{Number of correctly classified positive, negative, and neutral reviews}}{\text{Number of manually classified relevant positive, negative, and neutral reviews}}$$

$$\text{Precision} = \frac{\text{Number of correctly classified positive, negative, or neutral reviews}}{\text{Number of automatically classified positive, negative, or neutral reviews}}$$

$$\text{Recall} = \frac{\text{Number of correctly classified positive, negative, or neutral reviews}}{\text{Number of manually classified relevant positive, negative, or neutral reviews}}$$

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Figure 5. Confusion Matrix Metrics Formulas

An important note is that while F1 score is similar to accuracy, it also takes into account uneven class distribution, which allows it to be more helpful in measuring success versus accuracy. While calculating these scores, two different averaging methods - micro and macro averaging - were utilized in order to take into account that multilabel (non-binary) classification was being used to predict three classes: Positive, Negative, and Neutral. Micro averaging counts the total number of TP, FN, and FP, while macro averaging finds the unweighted mean for each label without taking the imbalance of labels into account. Table 9 shows these metrics split by polarity and averaging method.

Table 9.
CLASSIFICATION PERFORMANCE:
By Polarity / Averaging

	Sentiment Metric		
Class / Averaging Method	Precision	Recall	F1 Score
Negative	0.62	0.61	0.61
Neutral	0.57	0.87	0.69
Positive	0.90	0.72	0.80
Micro Average	0.73	0.73	0.73
Macro Average	0.70	0.74	0.70
Weighted Average	0.77	0.73	0.74

Note. Disparities in scores among classes may be due to sample sizes.