**Team 4 Project Final Report**

Samantha Wang, Mohammad Shafique, Kristina Sarkissyan, Melis Ocal, Nicholas Exel Moreira de Andrade, José J. Cao-Alvira, and Chris Angelidis

**Introduction and Objective**

The pharmaceutical industry generated 1.27 trillion dollars in revenue during 2020.[[1]](#footnote-0) It primarily concerns itself with the creation of medications, including R&D, manufacturing, and distribution. One of the key determinants of a medication’s efficacy is the patient's satisfaction regarding their treatment. By better understanding the factors associated with patient satisfaction, pharmaceutical companies can create more targeted and effective drug treatments. This can help patients recover faster and help the industry generate more revenue. With these goals in mind, our project has two main research objectives. The first is to build a model that can predict drug satisfaction based on the sentiment towards aspects found in the drug reviews, by using aspect based opinion mining to derive independent variables. The second is to ascertain which of these independent variables are the most impactful in determining drug satisfaction by using various traditional machine learning techniques.

**Data Description**

The source of the data is the Machine Learning Repository of the University of California at Irvine.[[2]](#footnote-1) The data is composed of 215,063 observations on patients reviews from the usage of specific drugs. For each observation, there is a review identifier, the name of the drug used, the condition treated by the drug, the patient’s review on the drug, a user satisfaction score on the drug (which assumes an integer value between one and ten), the date of the review, and the number of users who found the review useful. The reviews were amassed from April 2008 until September 2017 and were originally extracted from Drugs.com and Druglib.com, which independently collect information on drug usage and experience (Grässer, et al., 2018).

Chart, box and whisker chart

Description automatically generated

***Figure 1:*** *Depiction of satisfied vs. unsatisfied reviews in the dataset.*

We follow Grässer, et al. (2018) on our analysis and interpret the quantitative ratings in the data according to three levels of effectiveness; that is, assume that a rating greater than or equal to 7 is a positive or “satisfied” review, and a rating less than or equal to 4 is a negative or “dissatisfied” review. A review of 5 or 6 is considered to be neutral and will not provide enough sentiment to help in our analysis. A rating below 7 is considered “not satisfied”. Figure 1 contains a boxplot with the reviews according to the “satisfied” and “not satisfied” classifications.

**Preprocessing**

Standard preprocessing methods were employed on the text data to improve the quality of the results from the analysis. We normalized the medical conditions and reviews by making them lowercase. We removed quotes surrounding the reviews and fixed recurring errors that appeared in place of apostrophes; as we observed on multiple occasions, the review column contained the text “&#039” in place of an apostrophe. We additionally removed numbers, punctuation, and any irrelevant non-alphanumeric characters from the reviews. We tokenized each review and removed English stop words. Then, we incorporated the 20,000 most common words from Google[[3]](#footnote-2) and removed these words from the names of the drugs to prevent any discrepancies. We observed that there are corpus-specific stop words which we found in our initial analysis from the reviews and, in order to prevent any additional skew, we remove them as well.

**Methodology: Feature Extraction, Data Analysis & Text Analytical Methods**

For our initial feature extraction we defined the document features and created our test and training sets. We identified the 2000 most common words found in “satisfied” reviews (rating ≥ 7) and “dissatisfied” reviews (rating ≤ 4). Table A presents the 20 most common words associated with each of these classifications. It is readily evident that there is significant overlap between the two lists. There are only 2,314 unique words among both sets of 2000 words. We then used a Naive Bayes Classifier to provide us with an initial understanding of the data we are working with and its most informative features. For our Naive Bayes Classifier, our dependent variable is the classification of the reviews (satisfied or not satisfied) and our independent variables correspond to whether the review contains each of the 2314 unique words obtained from the two sets of 2000 most common words on “satisfied” and "dissatisfied" reviews. The Naïve Bayes is then trained on 90% of the data.

| **Satisfied** | **Dissatisfied** |
| --- | --- |
| effect | taking |
| side | pain |
| year | pill |
| take | month |
| taking | like |
| pain | effect |
| get | started |
| work | get |
| started | would |
| like | day |
| month | side |
| day | period |
| feel | took |
| pill | take |
| period | doctor |
| would | never |
| medication | feel |
| life | got |
| doctor | medication |
| week | week |

***Table A:*** *20 Most Common Words*

Panel I of Table B contains the most informative features extracted from training the classifier, and Panel II of the same table presents the performance statistics of the classifier on the remaining 10% of the data (ie. the test set). The feature with the highest likelihood of belonging to a “satisfied” review is containing the word “pleased”. Conversely, containing the word “lo” is the feature with the highest likelihood to be associated with a “not satisfied” review. It is important to note that the word “lo” likely appears mainly in a low dose birth control drug, which is typically how they are named. The performance statistics of the Naïve Bayes classifier show precision being the highest metric and accuracy the lowest. Based on this, we consider that the classifier may be biased towards the correct identification of “satisfied” reviews at the expense of not correctly identifying “not satisfied” reviews.

| **Panel I** | | |
| --- | --- | --- |
| **Features, contains:** | **Ratio** | **Likelihood** |
| lo | False : True | 12.3 : 1.0 |
| pleased | True : False | 10.2 : 1.0 |
| discontinued | False : True | 9.7 : 1.0 |
| poison | False : True | 9.7 : 1.0 |
| rubbish | False : True | 9.7 : 1.0 |
| tightness | False : True | 9.7 : 1.0 |
| vagina | False : True | 9.1 : 1.0 |
| confusion | False : True | 8.6 : 1.0 |
| poor | False : True | 8.1 : 1.0 |
| chose | True : False | 7.7 : 1.0 |
| easier | True : False | 7.3 : 1.0 |
| crap | False : True | 7.1 : 1.0 |
| labor | False : True | 7.1 : 1.0 |
| testosterone | False : True | 7.1 : 1.0 |
| flare | True : False | 7.1 : 1.0 |
| **Panel II**  **Performance Metrics** | | |
| **Accuracy** |  | 0.7300000 |
| **Precision** |  | 0.8088239 |
| **Recall** |  | 0.7971014 |
| **F-Score** |  | 0.8029197 |

***Table B:*** *Naive Bayes Model*

Additionally, we ran a logistic regression of the data where the dependent variable is the reviews classification (“satisfied” or “not satisfied”) and our independent variable is a weighted bag of words, employing a TFIDF vectorizer. The training dataset was composed of 90% of the data. Panel I of Table C contains the most informative features extracted from training the classifier, and Panel II of the same table presents the performance statistics of the classifier on the remaining 10% of the data. The feature with the highest likelihood of belonging to a “satisfied” review is containing the word “love”. Conversely, containing the word “worse” is the feature with the highest likelihood to be associated with a “not satisfied” review.

| **Panel I** | | |
| --- | --- | --- |
| **Features, contains:** | **Ratio** | **Likelihood** |
| love | True : False | 16.2 : 1.0 |
| miracle | True : False | 13.4 : 1.0 |
| amazing | True : False | 12.8 : 1.0 |
| saved | True : False | 12.1 : 1.0 |
| worse | False : True | 11.5 : 1.0 |
| changed life | True : False | 11.2 : 1.0 |
| disappointed | False : True | 10.4 : 1.0 |
| great | True : False | 9.8 : 1.0 |
| wonderful | True : False | 9.3 : 1.0 |
| lifesaver | True : False | 8.9 : 1.0 |
| year | True : False | 8.7 : 1.0 |
| worst | False : True | 8.5 : 1.0 |
| would love | False : True | 7.8 : 1.0 |
| useless | False : True | 7.3 : 1.0 |
| waste | False : True | 7.2 : 1.0 |
| removed | False : True | 7.1 : 1.0 |
| horrible | False : True | 6.6 : 1.0 |
| never recommend | False : True | 6.4 : 1.0 |
| depressed | False : True | 6.3 : 1.0 |
| **Panel II**  **Performance Metrics** | | |
| **Accuracy** |  | 0.94722648 |
| **Precision** |  | 0.94719682 |
| **Recall** |  | 0.94722648 |
| **F-Score** |  | 0.94721099 |

***Table C:*** *Logistic Regression Model*

We follow by calculating the Flesch Kincaid score for each review, Grade = 0.39\*µ+11.8\*π-15.59 where µ is total words by total sentences and π is total syllables by word. The Flesch Kincaid score is a widely used equation that assesses the reading level of a text. The scores range from 0 - 100. Lower scores correspond to a higher reading level. For example, a score ranging from 0 - 30 reflects college level writing.[[4]](#footnote-3)

For the next step, we employed a count vectorizer, where we chose the parameters min\_df and max\_df to eliminate corpus specific stopwords that appear over 25,000 times, as well as medical terms that only appear a few times. We then used Aspect-Based Opinion Mining to derive additional independent variables from the written text reviews. This technique has two parts, first identify *n* commonly occurring aspects amongst all reviews, and second, to determine what the reviewer sentiment is towards a given aspect. In order to identify the aspects, we used unsupervised learning methods leveraging on the Latent Dirichlet Allocation model (LDA). The input of this model is a word vector of frequency counts. The output of this model is a probability score that a review belongs to a given topic. In order to quantify sentiment, we use a rule-based sentiment score approach leveraging the VADER algorithm. At a high level, this algorithm assigns numerical scores to individual words in a vocabulary between -1 and 1. The more negative a word’s score is, the more negative the sentiment associated with the word. Similarly, the more positive a word’s score is, the more positive the sentiment associated with the word. A score of 0 is associated with a neutral sentiment. By running text through this algorithm, we are returned a compound polarity score which summarizes the overall sentiment score when considering all of the words in the given text. Ultimately by considering the sentiment score of each review as well as the probability that the review belongs to a given topic, we derive *n* independent variables. These are calculated as follows:

Topic N Probabilistic Compound Polarity Score = Probability of a Review Belonging to Topic N × Review Compound Polarity Score

The refinement of the count vectorizer contributed to effectively deriving greater meaning from our LDA aspects. For example, by removing additional stop words, it became easier to decipher what each topic was referring to. For completeness, we experimented with the count vectorizer using unigrams, bigrams, and the combination of unigrams and bigrams but did not find any benefit to utilizing bigrams or the combined unigrams and bigrams. As such, we decided to use the default range of only unigrams.

For our topic discovery, rather than using an algorithm to infer the topic names we decided to use our own judgement and analyzed the words to infer the subject of the topic and chose an appropriate name to match each topic.[[5]](#footnote-4) During our optimization process, we concluded on five topics to derive the most meaningful aspects. When we used fewer topics, side effects became a category as a whole instead of identifying *which* side effects we were attempting to refer to. When we selected more than five topics, we noticed that the boundaries between topics began to overlap and it became difficult to distinguish their meanings. The five topics identified are i. Skin Health, ii. Pain, iii. Gastro Health, iv. Menstrual Health and v. Mental Health. Figure 2 portrays the average VADER compound polarity score for each of these five topics. Reviews associated with Skin Health have on average a highly positive sentiment. The opposite is true for reviews associated with Pain.

Chart

Description automatically generated

***Figure 2:*** *LDA Topics and Vader Score Comparison*

**Model Evaluation**

Our estimation analysis is focused on each of the unique drugs included in the data set. For this, we aggregate the reviews with respect to each drug, and the dependent variable of our models corresponds to whether the average drug review is classified as “satisfied” or “not satisfied”. Drug satisfaction is considered to be correlated with the medical condition treated by the drug, the users’ sentiment on specific health topics -as revealed by their reviews- and the readability of the users’ reviews. Conditions treated are represented by 50 dummy variables each one corresponding to a specified medical condition. During our testing and optimization process, we found that those 50 most relevant medical conditions represent 76.6% of all conditions. Hence, each one was assigned as an unique dummy variable, while the remaining ones were aggregated as “other”. The user’s sentiment on specific health topics per drug is obtained from the average value for each drug of the Topic N Probabilistic Compound Polarity Score. The readability of the user’s reviews on a drug is the average value of the Flesch-Kincaid grade score of the drug’s reviews.

Figure 3 is a scatter plot where reviews are aggregated by condition and are scattered according to their average rating and Polarity Score. As expected, there is a positive correlation between ratings and the Polarity Score. It is interesting to observe that Pain, the condition associated with most reviews (as per the size of the bubble), does not show this positive correlation. Birth control is the second most reviewed condition on our data set and has a low review score and a low Polarity Score. Figure 4 is a scatter plot illustrating the average rating and Polarity Score for each unique drug in the data. The size of each bubble is the number of Reviews. The reviews are more evenly distributed by drug than by topic.

Chart, bubble chart

Description automatically generated

***Figure 3:*** *Drug Rating and Vader Compound Polarity Score per Condition & Number of Reviews broken down by disease being treated*

Chart, scatter chart

Description automatically generated

***Figure 4:*** *Drug Rating and Vader Compound Polarity Score per Condition & Number of Reviews broken down by individual drug*

Drug satisfaction is first estimated using the previously mentioned attributes employing a decision tree, a random forest, and a boosted tree classification method. A grid search method was used to identify the set of parameters that optimized the precision of each of the models. To further improve the precision of the estimation, we construct two ensemble models, stacking and voting, built off our three optimized models. Our models are trained based on 70% of the analyzed data, and performance metrics are calculated based on the remaining 30% of the data. Tables D and E summarize the results of the model evaluations. As indicated in Table D, all performance metrics of the Random Forest model are superior to the other two models. Table E presents the specification and performance statistics for the two ensemble models. The stacking model uses a logistic regression as the final estimator. The stacking ensemble outperforms the voting ensemble on all performance metrics. It is interesting to note that the stacking ensemble model outperforms the optimized Random Forest (the best performing single model) on all metrics except Recall.

| **Classification Models** | **Model #1** | **Model #2** | **Model #3** |
| --- | --- | --- | --- |
| *Decision Tree* | *Random Forest* | *Boosted Trees* |
| **Optimized**  **Parameters** | Criterion:  Entropy  Max Depth:  10 | Criterion:  Entropy  Estimators:  100  Bootstrap:  True | Loss:  cExponential  Estimators:  300 |
| **Accuracy** | 0.688 | **0.749** | 0.719 |
| **Precision** | 0.751 | **0.793** | 0.771 |
| **Recall** | 0.768 | **0.824** | 0.800 |
| **F-Score** | 0.759 | **0.808** | 0.785 |

***Table D:*** *Classification Models*

| **Ensemble Models** | **Model #1** | **Model #2** |
| --- | --- | --- |
| *Stacking* | *Voting* |
| **Optimized Classification**  **Models** | Decision Tree  Random Forest  Boosted Trees | Decision Tree  Random Forest  Boosted Trees |
| **Final Estimator** | Logistic Regression  {penalty= 'l2',  solver= 'newton-cg'} | NA |
| **Accuracy** | **0.750** | 0.733 |
| **Precision** | **0.796** | 0.778 |
| **Recall** | **0.820** | 0.817 |
| **F-Score** | **0.808** | 0.797 |

***Table E:*** *Ensemble Models*

Figure 5 presents the importances of the main attributes of the optimized individual classification models and the ensemble stacking model. The users’ sentiment towards the Gastro Health topic, proved to be the most important attribute on drug satisfaction. Following Gastro Health, was the user's sentiment towards the other four topics we identified, Pain, Skin Health, Mental Health, and Menstrual Health. The readability of a drug’s review carries additional importance on the estimation of satisfaction. Among health conditions treated, birth control appears to be the most important of the identified conditions affecting drug satisfaction. This is an interesting feature as birth control is one of the most reviewed conditions in our data set and is on average lowly rated by the users.

Chart, bar chart

Description automatically generated

***Figure 5 -*** Attribute Importances of the optimized individual classification models and the ensemble stacking model

As a basis of comparison, we also ran an Ordinary Least Squares Regression with the same independent variables as our previous classifications models but now the dependent variable is a continuous variable representing the average review score of each drug. Table F contains the highest 15 estimated coefficients for the regressors and the performance statistics. It is noteworthy that drug satisfaction with Menstrual Health has the highest estimated coefficient of all regressors.

| **Regressor** | **Coefficient** |
| --- | --- |
| Menstrual Health | 3.33 |
| Skin Health | 2.72 |
| Gastro Health | 2.49 |
| Mental Health | 2.03 |
| condition\_bowel preparation | -2.03 |
| condition\_overactive bladde | -1.87 |
| condition\_abnormal uterine bleeding | -1.80 |
| condition\_panic disorde | 1.78 |
| condition\_birth control | -1.68 |
| Pain | 1.62 |
| condition\_vaginal yeast infection | -1.03 |
| condition\_emergency contraception | 0.93 |
| condition\_anxiety and stress | 0.88 |
| condition\_back pain | 0.85 |
| condition\_diabetes, type 2 | -0.71 |
| **Mean Error (ME)** | 0 |
| **Root Mean Squared Error (RMSE)** | 1.8716 |
| **Mean Absolute Error (MAE)** | 1.3994 |
| **Mean Percentage Error (MPE)** | -19.4793 |
| **Mean Absolute Percentage Error (MAPE)** | 34.8552 |

***Table F:*** *OLS Regression*

For robustness, we expanded our analysis and included data on the user’s cost of each drug with the intention to analyze the effects pricing may have on drug satisfaction and verify if our previous findings still hold after adding this variable. The source of the data is the Centers for Medicare and Medicaid Services and includes information on household spending on drugs paid through the Medicaid program. For each drug observation there is Brand Name and Spending, which is the average cost per dose of the drug. For preprocessing the pricing dataset, we normalized the drug names similarly to how we normalized them for our reviews dataset; that is by lowercasing the names and removing any unnecessary non-alphanumeric characters. We also noticed that the drug names on the pricing dataset had a different format when referring to multiple drugs, so we replaced the “-” with “/” in the pricing dataset in order to better match the drug names of the reviews dataset.

Figure 6 depicts a scatter plot with the average rating on the horizontal axis and the VADER score on the vertical axis for each drug, where the size of each bubble is the average household spending on a drug. Not surprisingly, the bigger bubbles are on the bottom-left quartile of the graph. Drug satisfaction is estimated following the optimization of classification methods and ensemble models as previously described, but now incorporating the data on the costs of each drug. Tables G, Table H, and Figure 7 summarize the results of the models’ evaluations.

Chart, bubble chart

Description automatically generated

***Figure 6:*** *Drug Rating and Vader Compound Polarity Score per Drug & Spending*

As in the previous iteration of models, the performance of the Random Forest model is superior to that of the other two considered models and the ensemble stacking model produces better performance on all metrics than the voting model. Yet, the original model estimation using ensemble stacking, (i.e. not including the drug cost data), produces better performance on two statistics (Precision and Recall) compared to the model including pricing data. The Accuracy and F-score on both models are almost identical.

Figure 7 presents the importances of the main attributes of the optimized individual classification models and the ensemble stacking model when adding pricing data. The users’ sentiment towards Skin Health now proves to be the most important attribute on drug satisfaction. The analysis also indicates that while readability of review continues being important, spending on drug treatments is now more important.

| **Classification Models** | **Model #1** | **Model #2** | **Model #3** |
| --- | --- | --- | --- |
| *Decision Tree* | *Random Forest* | *Boosted Trees* |
| **Optimized**  **Parameters** | Criterion:  Entropy  Max Depth:  4 | Criterion:  Entropy  Estimators:  200  Bootstrap:  True | Loss:  Exponential  Estimators:  100 |
| **Accuracy** | 0.676 | **0.747** | 0.727 |
| **Precision** | 0.727 | **0.764** | 0.752 |
| **Recall** | 0.744 | **0.844** | 0.820 |
| **F-Score** | 0.735 | **0.802** | 0.785 |

***Table G:*** *Classification Models with Drug Spending*

| **Ensemble Models** | **Model #1** | **Model #2** |
| --- | --- | --- |
| **Type** | *Stacking* | *Voting* |
| **Optimized Classification**  **Models** | Decision Tree  Random Forest  Boosted Trees | Decision Tree  Random Forest  Boosted Trees |
| **Final Estimator** | Logistic Regression  {penalty= 'l2',  solver= 'newton-cg'} | NA |
| **Accuracy** | **0.752** | 0.733 |
| **Precision** | **0.774** | 0.761 |
| **Recall** | **0.834** | 0.815 |
| **F-Score** | **0.803** | 0.787 |

***Table H - Ensemble Models with Drug Spending***

Chart, bar chart

Description automatically generated

***Figure 7:*** *Importance of main attributes after incorporating the pricing data.*

**Conclusions**

Our analysis uncovered that drug satisfaction is highly associated with user sentiment with the five predominant topics present in the reviews; these are i. Skin Health, ii. Pain, iii. Gastro Health, iv. Menstrual Health and v. Mental Health. This holds true even after controlling for spending on drug treatments. Drug satisfaction with Gastro Health is the most important topic across various analyses. In addition, satisfaction on Skin Health, Menstrual Health, Pain and Mental Health proved to be just as important.

**Practical implications**

The presented analysis, models, and results provide insight into the practical implications of our research. We are able to identify key text features that can be used to identify if a review is good or bad. We are able to provide a more insightful and replicable metric for measuring drug satisfaction as compared to a traditional 1-10 scale. Doctors can harness the experience of patients who have previously used a given drug and if the drug proves to have negative sentiment, current patients can be provided with clear feedback on the efficacy of the drug. Using our findings to improve customer sentiment, as well as including drug pricing, can be useful for health coverage providers in helping them determine which drugs to provide coverage for. Lastly, this can also be useful for pharmaceutical companies in helping them further develop and improve their drugs.

**References:**

Grässer, F., 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). ACM, New York, NY, USA, 121-125.

Mankad, S., Hyunjeong “Spring” Han, Joel Goh, Srinagesh Gavirneni (2016) Understanding Online Hotel Reviews Through Automated Text Analysis. Service Science 8(2):124-138.

CMS.gov. 2021. *Medicaid Drug Spending Dashboard | CMS*. [online] Available at: <<https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Information-on-Prescription-Drugs/Medicaid>>

Medium. 2021. *Flesch-Kincaid Grade Level: How hard is it?*. [online] Available at: <<https://medium.com/@annwylie/flesch-kincaid-grade-level-how-hard-is-it-ebb0bfcdfa87>>

Statista. 2021. *Topic: Global pharmaceutical industry*. [online] Available at: <<https://www.statista.com/topics/1764/global-pharmaceutical-industry/#dossierKeyfigures>>

1. <https://www.statista.com/topics/1764/global-pharmaceutical-industry/#dossierKeyfigures> [↑](#footnote-ref-0)
2. Dataset available online at:<https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29> [↑](#footnote-ref-1)
3. 20,000 Most Common Words: <https://github.com/first20hours/google-10000-english/blob/master/20k.txt> [↑](#footnote-ref-2)
4. <https://medium.com/@annwylie/flesch-kincaid-grade-level-how-hard-is-it-ebb0bfcdfa87> [↑](#footnote-ref-3)
5. When trying a uniform algorithm to infer what the topics were, we concluded that the outputs were not insightful. [↑](#footnote-ref-4)