Sentiment Analysis of Drug Review Data to Classify Positive or Negative Rating

Amin Fesharaki, Conor Fitzpatrick, Uyen Pham

University of San Diego

ADS 504: Machine Learning

August 15, 2022

Abstract

As advances in pharmaceutical technology expand, patients now have more options than ever for which type of medication they use for a particular ailment. While this has traditionally been decided by a healthcare provider, there is now a wealth of information online about real-world experiences with various prescription medications, accessible to anyone. This project analyzes a large text dataset of drug reviews, and applies sentiment analysis to determine whether the reviews are “positive” or “negative”. With numerous machine learning algorithms tested, the Decision Tree returned the best results, with an accuracy of 90% in classifying positive or negative reviews. These results are highly encouraging, and more work will no doubt reveal even greater sentiment analysis accuracy.

Table of Contents

Introduction

Patients now have more options than ever for prescription drugs, and with advertising and direct-to-consumer marketing at an all-time high, it can be confusing for patients to decide what drugs may work best for them (DeFrank et al., 2020). Certain drugs that treat the same condition may have different side effects, or some may have better results in certain subsets of people. With the widespread adoption of the internet, it is now possible for people to post their own experiences with prescription medications, providing a great resource for others in similar situations or seeking similar treatment. Being able to sift through large amounts of drug reviews to pick out certain features that may be relevant to an individual, therefore, would be a huge benefit to the healthcare realm. In addition to patients, drug companies and healthcare providers would benefit from a robust analysis of drugs and their effects and side effects in a real-world context. This project aims to sift through a large dataset of text-based reviews on numerous drugs and will attempt to predict the sentiment, either positive or negative, based solely on the review itself.

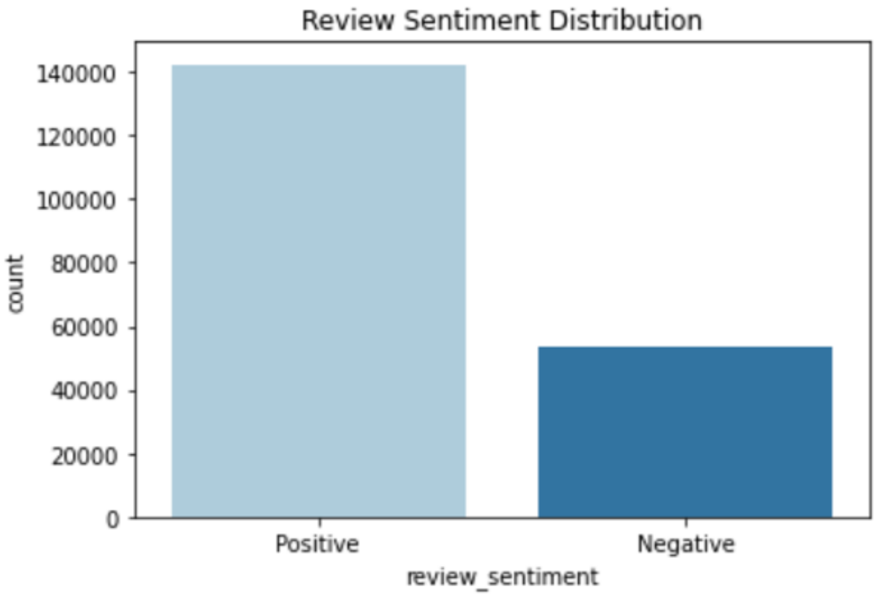
The data was downloaded from the UCI Machine Learning Repository, which was originally in two sets, a training (161,297 instances) and a test set (53,766 instances). These were combined into a dataset containing 215, 063 instances. Originally there were seven features, including ID, Drug name, Condition, Rating, Review, Date, and Usefulness Rating. This project focuses on a sentiment analysis of the text-based reviews to predict the rating.

Methodology

The data were combined into one set at the beginning for some text preprocessing. The original rating (1-10) showed imbalance classes with the most common rating being at the two extremes of the rating (1 and 8-10). The task purpose was set to focus on the extreme ratings, but not the neutral ones. Therefore, a binary class was generated with a scale of 7-10 considered positive and 1-4 as negative. The less interesting ratings of 5-6, were dropped. This approach could help make the huge data set less complicated. Figure 1 shows the distribution of positive and negative ratings. The two classes were still imbalanced, but not significant, so sampling was not necessary.

Figure 1

Positive and negative drug review distribution



The review text was cleaned to be free of capital letters, punctuations, special characters, ASCII characters, trailing white space, etc. using parsing technique. Also, The Natural Language

Toolkit (NLTK) was applied to remove commonly used words which did not contribute much to the classification task (Judah, 2021).

The dataset was split into training (80%) and test set (20%). Then a TF-IDF vectorizer method was performed on both sets separately. The vectorizer would generate a large parse data set, so setting the max-feature value to 5000 reduced the dimension of the data dramatically, thus reducing the run time and potentially helped to improve the performance of selected models.

Different models were applied. Some were with tuning and cross-validation to find the best parameter and then applied to the test set. The studied models include Decision Tree (DT), Logistic Regression (LR), Single Layer Perceptron, Support Vector Machine Learning (SVM), Logistic Regression with regularization, Naive Bayes, K-Nearest Neighbors, and ADABoost.

Accuracy was chosen to be the metric measuring the success of these models because the task was to focus on how well the models could correctly classify both classes. Negative and positive reviews are both important feedback for drug companies to evaluate their products as well as for consumers to be aware of the sentiment of the drug from other users.

Results and Final Model Selection

The first model tested was a Decision Tree. Decision Trees use recursive binary splitting, which branches decisions off from a root node into numerous splits. This model returned the best accuracy, at 89.8%. The maximum depth of the tree was tuned, with iterations ranging from 50 to 600, with 200 being the optimal maximum depth of the tree when balancing complexity and accuracy.

A Support Vector Machine (SVM), which utilizes hyperplanes in a high dimensional space to perform classification, was run next with a hinge loss function. This returned an accuracy of 85% utilizing 5-fold cross-validation. A Single Layer Perceptron, which is a simple and early form of a neural network, was also run utilizing 5-fold cross-validation, and returned an accuracy of 80.7% on the test data. A logistic regression model with L1/L2 regularization returned a slightly better accuracy of 86%.

A Bernoulli Naive Bayes model was run on the data, which is based on the Bayes Theorem, which returned an accuracy of 82%. A simple K-nearest neighbor mode was then run, utilizing a K value of 3, which returned an accuracy of 74%. Finally, an ADA Boost model, which is short for Adaptive Boosting, returned an accuracy of 79%.

Table 1

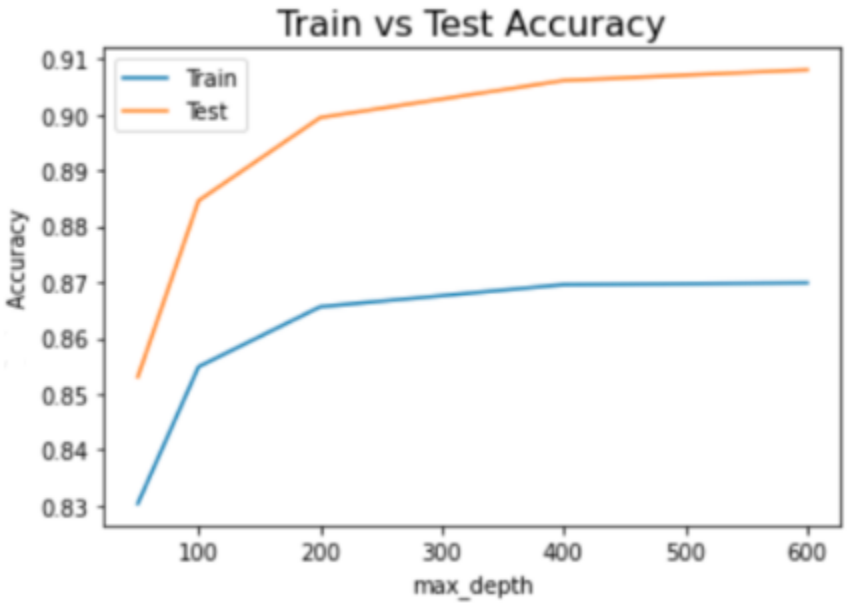
Model Evaluation Table for Sentiment Analysis

Evaluation Measure	**Decision Tree**	**Support Vector Machine**	**Perceptron**	**Logistic Regression with Regularization**	**Naive Bayes**	**KNN**	**ADA Boost**
Accuracy	**0.90**	0.85	0.82	0.86	0.82	0.74	0.79
Precision	**0.93**	0.86	0.85	0.88	0.83	0.85	0.78
Recall	0.94	**0.96**	0.91	0.93	0.82	0.74	0.79
F1 Measure	**0.93**	0.90	0.88	0.91	0.83	0.76	0.77

By comparing model metric performance in Table 1, the Decision Tree model outperformed all other models with an accuracy of 90%. In addition, the Decision Tree model classified the most positive reviews correctly with a precision of 93% which is the proportion of positive identifications that was actually correct, the second highest recall score (proportion of actual positives was identified correctly, and the highest F1: the harmonic mean of precision and recall. Moreover, the max tree depth was varied and compared to determine the optimal hyperparameters of the model. The accuracy rating over a max tree depth of 200 was deemed negligible and not worth the model complexity. Therefore a max depth of 200 was chosen when balancing complexity and accuracy. Figure 2 supports this notion as it illustrates the train and test accuracy when varying the tree depth from 50 to 600.

Figure 2

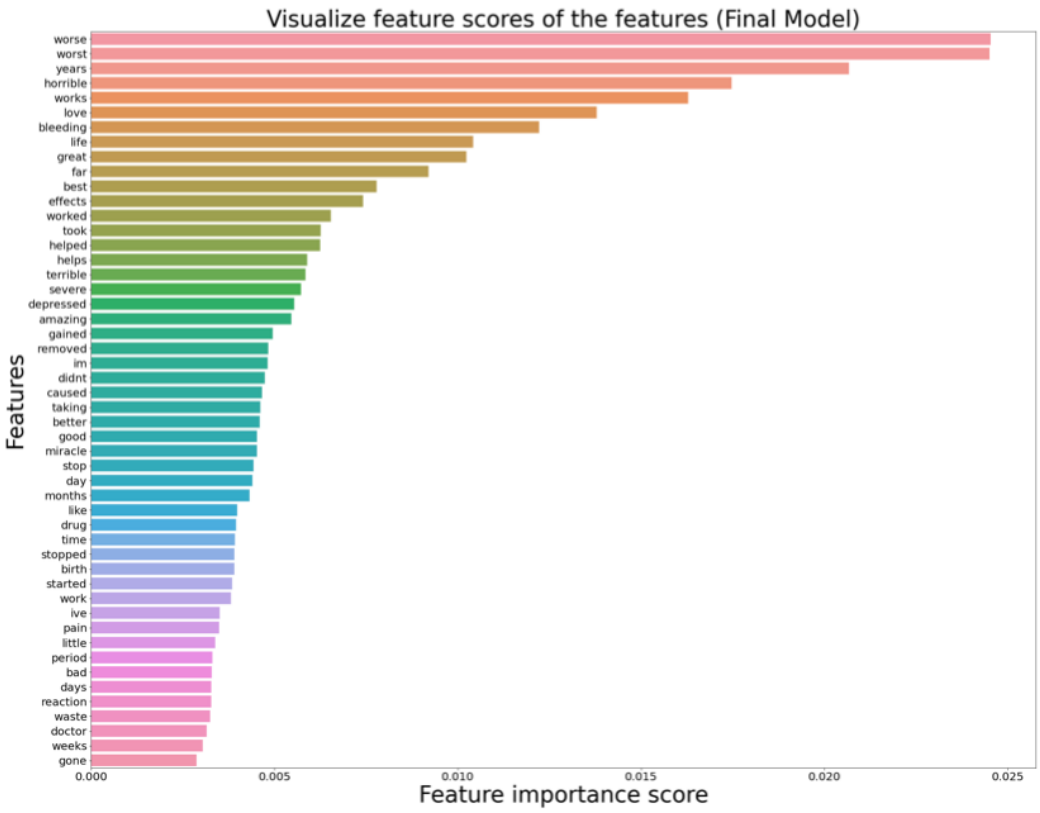
Train and test accuracy of the Decision Tree model when varying tree max depth



In addition, feature importance was included for the decision tree model to gain a better understanding of what words were deemed most important for the model. The top 50 words for the Decision Tree model with a max tree depth of 200 is shown in Figure 3. The majority words shown in the feature importance graph include “strong” words that correlate to negative and positive reviews. However, further analyzing the figure shows potential areas where the model can improve by eliminating and combining certain words and phrases during feature transformation. For example, the top 2 words, “worse” and “worst” can be combined into a singular phrase which could potentially improve model performance.

Figure 3

Feature importance for the Decision Tree model with a max tree depth of 200



Discussion and Conclusion

The healthcare industry can utilize machine learning in many different ways to improve patient care by analyzing vast amounts of medical records and patient data beyond the scope of human capability. There are many ways drug companies and healthcare providers can benefit from machine learning's ability to gather insights through algorithms. In this study, 150,000 drug-related text reviews are analyzed to predict the positive or negative sentiment which can be used for improving drugs, market research, customer service, or brand and producing monitoring.

Furthermore, the study incorporated eight classification models that were used to classify whether a review was "positive" or "negative". In conclusion, the Decision Tree model outperformed all other models in accuracy, precision, and F1 score as well as having the second highest recall score. With an accuracy of 90%, the Decision Tree model is considered successful enough to be implemented as a tool to analyze a binary sentiment analysis on drug text reviews.

Future investigations can delve deeper into specific word and phrase combinations to gain a more granular understanding of the sentiment specific to drug reviews, which may lead to an increase of accuracy. In addition, a multiclass classification can also be investigated to provide a more accurate description of reviews (i.e, very negative, negative, neutral, positive, very positive), and ultimately advance to a 1-10 rating system. By doing so, a greater in-depth understanding can be achieved to better improve drug quality and increase drug company revenue.

References

Judah, B. (2021, November 10). Removing stop words with NLTK library in Python. *Analytics Vidhya*. https://medium.com/analytics-vidhya/removing-stop-words-with-nltk-library-in-python-f33f53556cc1

DeFrank, J. T., Berkman, N. D., Kahwati, L., Cullen, K., Aikin, K. J., & Sullivan, H. W. (2020). Direct-to-Consumer Advertising of Prescription Drugs and the Patient–Prescriber Encounter: A Systematic Review. *Health Communication*, *35*(6), 739–746. https://doi.org/10.1080/10410236.2019.1584781

Appendix