ReddPill: A machine-learning framework for analysing and predicting user attitudes toward medicines using the social media platform Reddit

Author: James Mathews

Date: 04-Jan-2022

Course: Harvard edX - Data Science Professional Certificate

Project: Capstone - Choose Your Own

Introduction

Data scraping and machine learning techniques offer innovative new ways to understand public attitudes toward medicines. A 2022 study by Spadaro et al. extracted data from the social media website Reddit and analysed this data to provide insights into current trends in opiate addiction (Spadaro et al. 2022). This current project planned to build on these ideas, investigating the feasibility of using machine learning techniques to understand and predict user attitudes toward a medicine on the Reddit platform.

As a test case, the drug pregabalin (originally patented as Lyrica) was chosen. Pregabalin is a GABA analogue used as an anticonvulsant, anxiolytic, analgesic, and off-label in other indications (Abai 2019). Due to its atypical mechanism of action and relatively recent introduction to the market, questions still remain as to public attitudes toward and recreational use of the drug. Insights gathered by this analysis could suggest opportunities for public health interventions, physician education, or targeted digital campaigns.

The report was compiled using R Markdown in RStudio, an integrated development environment for programming in R, a language and software environment for statistical computing (R Core Team 2021).

Methods - Part 1

Data scraping

Generating a significantly sized dataset from Reddit proved to be a great obstacle in this project. To this end, several approaches were explored. The first attempt invoked functions from the minimalist Reddit scraping package RedditExtractoR by Ivan Rivera (Rivera 2022). RedditExtractoR makes use of Reddit's native App Program Interface (API) to extract and parse formatted data directly into R. Though convenient, this approach is constrained by the Reddit API's configuration, which imposes a limit to the number of posts that can be retrieved. A dataset consisting of 240 Reddit submissions sorted by top rating from the previous year was extracted in this manner. Because of the above mentioned API constraints, user data could not be extracted concurrently in an automated manner using RedditExtractoR.

Alternative data scraping methods were explored so as to generate a more appropriate dataset for analysis. Spedaro et al.'s data extraction used the Python Reddit API Wrapper (PRAW) tool to extract more than 2*10^5 reddit posts, and so these methods were replicated. Unfortunately, since Spedaro's work, the Reddit API has been modified to dissallow iterative CloudSearch based requests that might yield a substantial dataset.

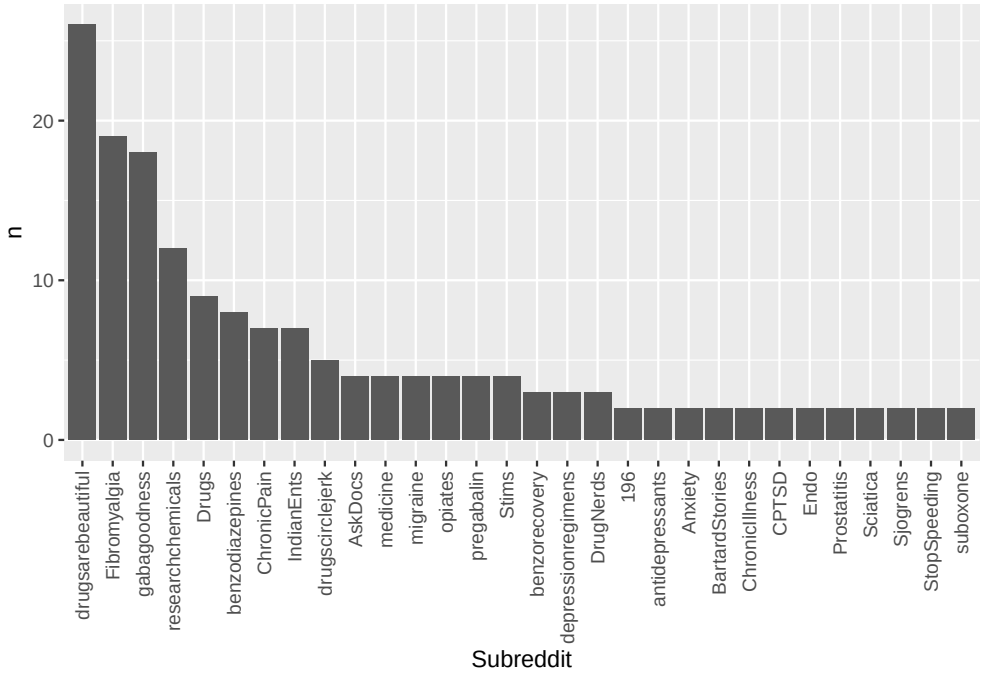
Another option for data scraping relied on the third party Reddit archiving service Pushshift.io, maintained by the Reddit user /u/stuck_in_the_matrix (Baumgartner et al. 2020). Pushshift.io offers an alternative API to extract potentially unlimited Reddit content. The PRAW based approach mentioned above was modified to direct toward the Pushshift.io API rather than Reddit and configured to download `r1*10^4‘ submissions from Reddit. However, due to instability with Pushshift.io's API, attempts to extract data in this manner were not successful. An R based approach to access Pushshift.io's API was also tested tested without success (commented out in the code of this report). A copy of the Python script mentioned is avaialble from this writer's Github account. An example dataset produced from RedditExtractor is also available at Github.

Exploratory analysis

The small dataset that could be extracted was analysed for insights and to select a suitable machine learning approach. This dataset consisted of 240 rows describing the variables date_utc, timestamp, title, text, subreddit, comments, url, date, text_clean, sent. Each row in the dataset represents one submission (comments on other submissions are excluded) to any board on Reddit that includes the word "pregabalin" in the text or title. Number of comments on each post was a variable included in final analysis but not explored here.

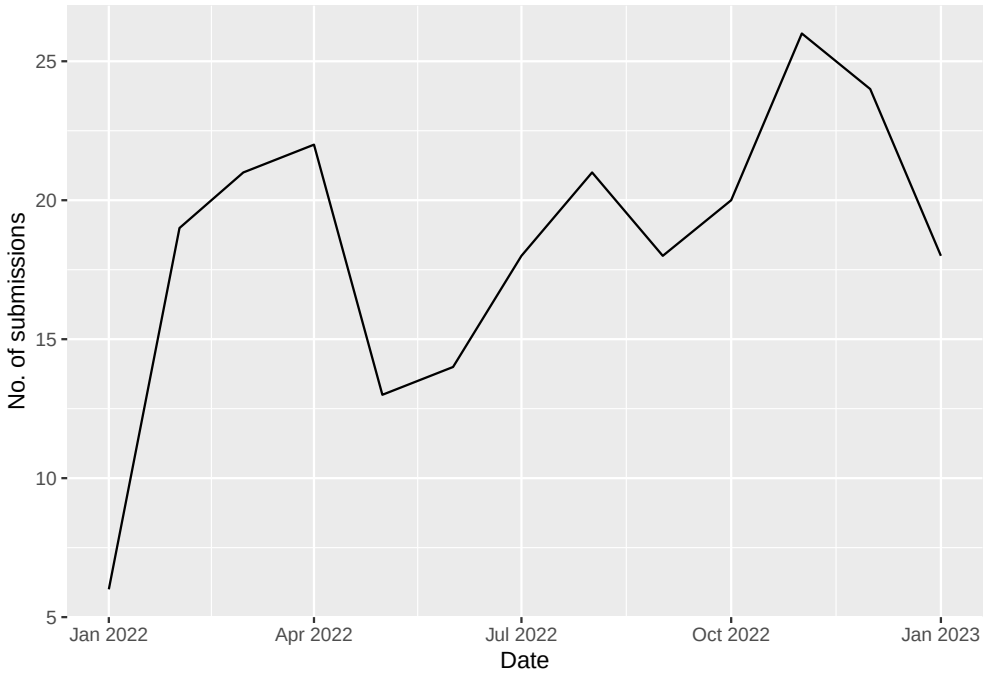
Exploratory analysis: Subreddit

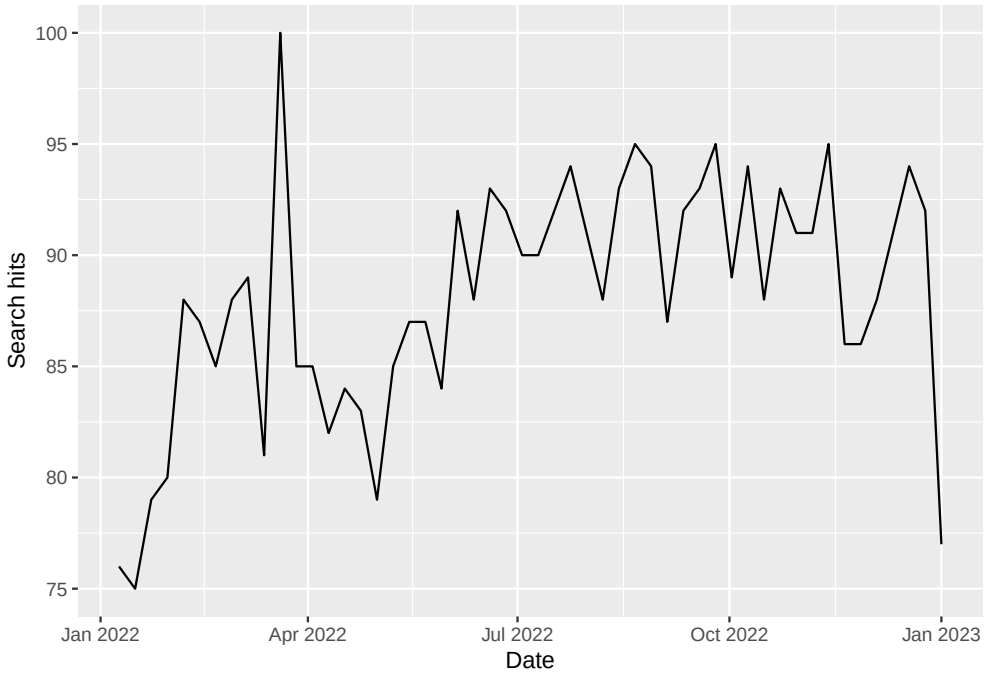
Reddit is organised into many different messageboards called subreddits. Each subreddit is a community dedicated to discussion of a particular topic. Figure 1 shows the 25 most common subreddits found in the dataset, with the most popular being drugsarebeautiful. Already, it is clear from the nature of these subreddits that a majority of pregabalin discussions on the platform relate to recreational use and addiction. A minority relate to discussion of legitimate medical use.



Exploratory analysis: Date

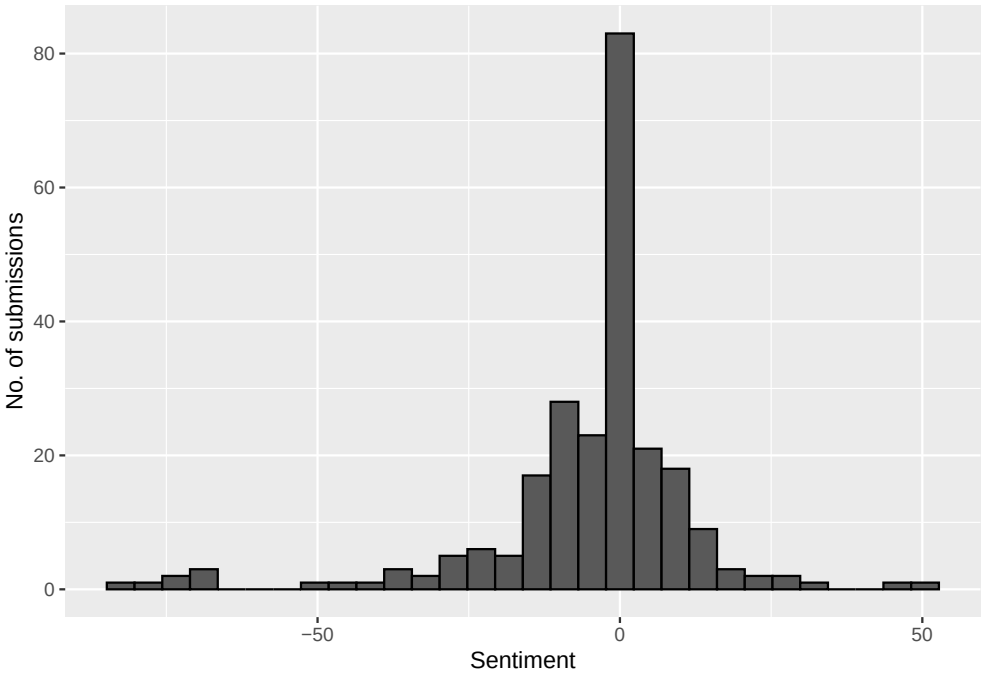
The "date_utc" column in the scraped dataset represents the date on which each submission was originally made. In order to effectively visualise this information, date was rounded to the nearest week using the lubridate package and frequencies plotted (Figure 2). Submissions from later in the year were more likely to be included in the dataset, and it is not entirely clear why. A cursory analysis of Google search data using the gtrendsR package (Figure 3) suggests that this may correlate with a global pattern of rising interest in pregabalin (Massicotte and Eddelbuettel 2022).





Exploratory analysis: Title and Text

The title and text fields of the dataset contain the writing of Redditors that will be used to drive machine learning techniques in this project. In order to make this information more amenable to analysis, preprocessing was performed with the tidytext package. First the title and text of each post was concatenated. HTML tags, numbers, and symbols were removed. Shortened words were then replaced and whitespace was trimmed. The resulting text had a mean length of 2316.9375 characters with a minimum of 6 and a maximum of 37642.



In order to understand user attitudes toward pregabalin, a sentiment score was calculated for the cleaned text of each post. The syuzhet package and the afinn sentiment library were used to this end and the distribution of sentiments by post plotted as a histogram (Figure 3) (Jockers 2015). The mean sentiment for all posts was -4.65, overwhelmingly negative. There were outlying posts with a very positive sentiment score (>50) or overwhelmingly negative score (<-50). Sentiment distribution broadly resembled a normal (Gaussian) distribution.

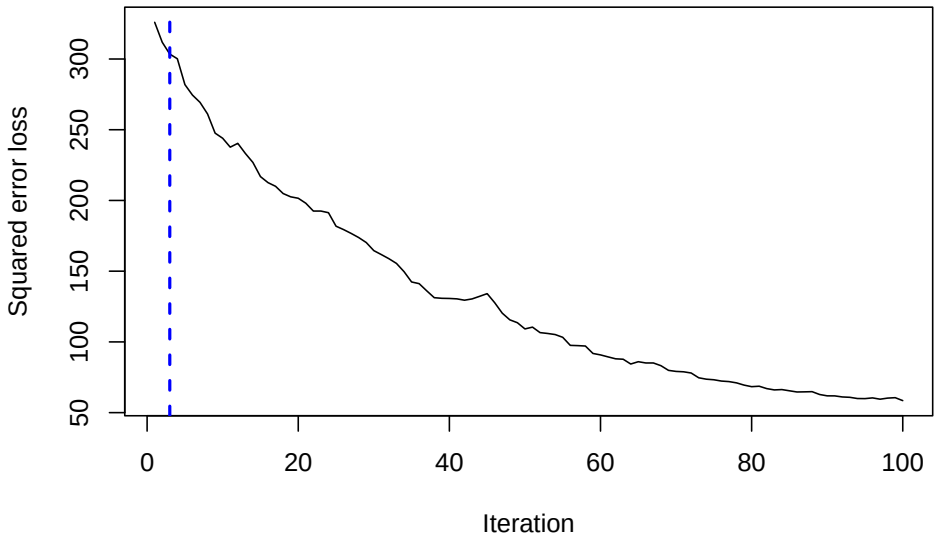
Methods - Part 2

Machine learning: GBM

Two different machine learning algorithms were implemented in an attempt to predict sentiment outcomes on the basis of other Reddit post characteristics. The first algorithm selected was Stochasic Gradient-Boosted Tree Modelling (GBM) using the gbm package (Friedman 2001). GBM aims to create a function that can predict the value of a variable by fitting a function over observations of that variable and iterating this function, each time adding further terms in an attempt to minimise error loss. GBM was chosen for this application because it is suitable to work with large depth factors (eg. subreddit) as predictors.

First, the dataset was partitioned into testing and training sets, with the training set consisting 80% of the data - a partition rate dictated by the small size of the overall dataset. A simple instance of gbm with default parameters was called after first converting character vectors to factors. Root Mean SquareError (RMSE) was used to measure model accuracy in all cases. The RMSE achieved with this approach was 347.0703985, indicating further optimisation was required.

On consulting documentation for the gbm package, a number of adjustments were made in order to optimise the algorithm. For prediction over a small sample size with few predictors, interaction depth and minimum numver of observations per node were both lowered. The gm.perf function was called to estimate the optimal number of trees (Figure 4). Then, a grid of different shrinkage (learning rate) values were tested.



Machine Learning: SVM

Support Vector Machines (SVMs) are similar to GBM models in that they are tree based supervised learning models that are suitable for classifying according to factor variables. As opposed to GBM, SVM models applying decision trees to directly categorise datapoints rather than fit a function to describe them. A simple SVM model was first attempted with default parameters using the e1071 package (Cortes and Vapnik 1995).

Results

Algorithm 1: Simple GBM

The simple GBM model developed was applied to the test dataset and an RMSE of 15.8079131 was calculated. This indicated a very poorly algorithm, and so steps were taken to optimise the model.

Method	RMSE
Simple GBM	15.80791

Algorithm 2: Optimised GBM

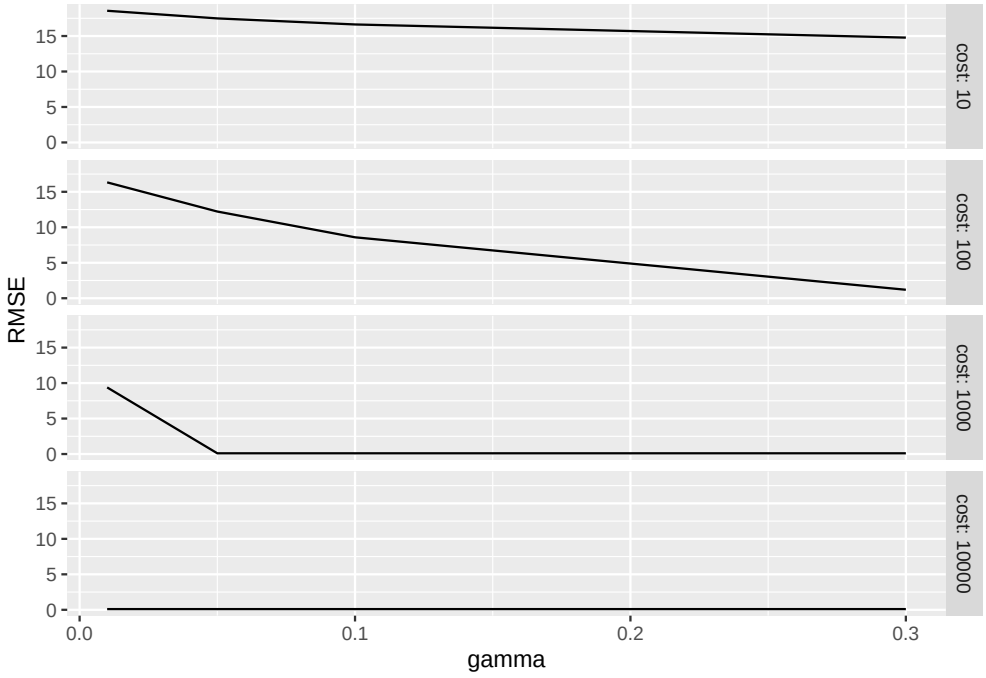
The final optimised GBM model used 347.0703985 trees and a learning rate of 0.005. When this algorithm was used to predict sentiment scores in the test dataset, an RMSE of `r` was achieved.

Method	RMSE
Simple GBM	15.8079131172198
Optimised GBM	15.911305842016

Algorithm 3: Simple SVM

As an alternative to GBM, an SVM model was also explored for its suitability to predict sentiment of posts. Before optimisation, this SVM model returned an RMSE of 16.531124. Further optimisation was clearly necessary.

Method	RMSE
Simple GBM	15.8079131172198
Optimised GBM	15.911305842016
Simple SVM	16.5311239966186



Algorithm 4: Optimised SVM model

Optimisation of the aforementioned SVM model was performed using a tuning grid of different values for the parameters cost and gamma (Figure 6). The optimal value for gamma was 0.01 and for cost was 10^4. When the optimised model was used to predict the sentiment of the test set, an RMSE of 19.46576 was achieved.

Method	RMSE
Simple GBM	15.8079131172198
Optimised GBM	15.911305842016
Simple SVM	16.5311239966186
Optimised SVM	19.4657599679711

Conclusion

This project aimed to extract data from the social networking site Reddit and use machine learning to contextually predict post sentiment toward a particular medicine, with the GABA analogue pregabalin taken as a test case. While it was easy to extract a small quantity of data from Reddit, efforts to scrape large datasets were confounded by restrictions on the site's API that limited how much data would be sent. Instability of the API hosted by third-party archiving service pushshift.io meant that this was not a viable alternative datasource. Nevertheless, analysis and machine learning were applied to a small dataset.

Analysis of Reddit discussions relating to the drug pregabalin were illuminating. It was clear that pregabalin was predominantly discussed in the context of recreational use and addiction. This fact may come as a surprise to prescribers, since the consensus in the medical literature is that pregabalin's abuse potential is relatively low (Papazisis and Tzachanis 2014). Perhaps unsurprisingly given the context of many pregabalin discussions, sentiment analysis of text revealed that users felt on average very negatively toward the drug. Interestingly, aggregations of posts were observed to have either a very positive or a very negative sentiment, suggesting that the Reddit community is polarised in its attitude.

Two different machine learning approaches were trialed in order to contextually predict the sentiment of Reddit posts that mentioned the drug pregabalin. A GBM algorithm was first trialed that, even after optimisation, showed fairly poor predictive power for this dataset. An SVM model was instead trialled and optimised, resulting in no significant improvement over GBM. Nevertheless, if more detailed realtime data from Reddit became available, the approaches tested in this project could be a powerful framework for understanding and predicting user attitudes to medicines.

References

Abai, B. 2019. *StatPearls*. StatPearls Publishing LLC. https://books.google.pl/books?id=xwxEzQEACAAJ.

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. "The Pushshift Reddit Dataset." *CoRR* abs/2001.08435. https://arxiv.org/abs/2001.08435.

Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20 (3): 273–97.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*, 1189–1232.

Jockers, Matthew L. 2015. *Syuzhet: Extract Sentiment and Plot Arcs from Text*. https://github.com/mjockers/syuzhet.

Massicotte, Philippe, and Dirk Eddelbuettel. 2022. *gtrendsR: Perform and Display Google Trends Queries*. https://CRAN.R-project.org/package=gtrendsR.

Papazisis, Georgios, and Dimitrios Tzachanis. 2014. "Pregabalin's Abuse Potential: A Mini Review Focusing on the Pharmacological Profile." *International Journal of Clinical Pharmacology and Therapeutics* 52 (8): 709—716. https://doi.org/10.5414/cp202118.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rivera, Ivan. 2022. *RedditExtractoR: Reddit Data Extraction Toolkit*. https://CRAN.R-project.org/package=RedditExtractoR.

Spadaro, Anthony, Abeed Sarker, Whitney Hogg-Bremer, Jennifer S. Love, Nicole O'Donnell, Lewis S. Nelson, and Jeanmarie Perrone. 2022. "Reddit Discussions about Buprenorphine Associated Precipitated Withdrawal in the Era of Fentanyl." *Clinical Toxicology* 60 (6): 694–701. https://doi.org/10.1080/15563650.2022.2032730.