Evidence Based Approach Using Multivariate Regression Analysis And Machine Learning Algorithms To Discover Effect of CBD on Arthritis Subjects

CONTRIBUTORS:
SETH KUTTY
CHRISTINA CHENEY
ABHIRAJ MALAPPA
XIJUN ZHANG
SHARATH KALKUR

Index Terms

Medical Cannabis; Arthritis; Osteo-Arthritis; Joint Pain; CBD; THC; Machine Language; Nutraceuticals; Pharmacology; Strain Efficacy; Strain side-effects

13 November 2017

Abstract

In the absence of randomized clinical trials, meta-analysis of crowdsourced consumer data may provide the most accurate understanding of how hemp-based CBD is currently used by patients in treating various medical conditions such as Arthritis. We compared different statistical and data science analytics approaches to determine how CBD [2],[3], (cannabidiol) is used to alleviate or relieve Arthritis pain. Two approaches, standard multivariate regression analysis and machine learning algorithmic models were used to classify, validate and predict CBD dosages for the specific medical condition of arthritis. Our analysis has identified CBD products that have the highest overall consumer ratings with the least side effects. Furthermore, we can infer from the data optimal CBD ranges where customers are experiencing high efficacies. Our results are summarized on the website

This research addresses the following issues for the beneficiaries:

- **Customers:** Arthritis support groups and focused patient groups would benefit from this analysis as it either validates their existing medical cannabis prescriptions or provides alternatives with the least side effects, without a trial-and-error approach to the plethora of strains available today.
- **Dispensaries:** Dispensaries can use these results to recommend appropriate CBD Products in their inventory for customers with Arthritis. They can also match their observational understandings with our study. E.g. if they do not carry a specific product with the CBD range that is recommended, they can look at similar products in the same category as an alternative.
- **Pharmacologists:** Dispensaries do not have any research around CBD, cannibanoid compounds and their effectiveness for specific medical conditions. This research paper provides them the required analysis to ensure their recommendations have a high percentage of efficacy.

1 INTRODUCTION

In the absence of conclusive randomized clinical trials, scientific analysis of non-experimental crowdsourced data can shed light on how patients are using CBD to treat specific medical conditions. This evidence based approach to crowdsourced

data with the right data science models applied against them will help define, classify and predict the efficacy of specific strains and compounds (i.e. CBD currently, and other non-psychoactive hemp-derived cannabinoids and terpenes in the future) and their relative side effects. Data from these observational studies can be pooled to provide an estimate of efficacy that may be more precise than that obtained by a single study.

Without sufficient clinical studies, CBD relies on anecdotal and trial-and-error approaches to cannabis strain recommendations and dosage without regard to side effects or patient constitution. Traditional aggregate data meta-analyses, especially of observational studies, have limited capacity to adequately account for factors because of the sparsity of the data from dispensaries, clinical practitioners and the lack of rigor in the patient feedback collection. While the crowdsourced data we collected contained information on patient usage for many medical conditions, we have narrowed our analysis to one condition here, specifically Arthritis.

1.1 Arthritis: An Introduction

More than 40 million people in the United States have doctor-diagnosed Arthritis, including almost 300,000 children. Osteo-Arthritis is the most common, affecting more than 21 million Americans. Rheumatoid Arthritis is the second most common condition, affecting approximately 2.1 million.

Arthritis refers to joint pain or joint disease and there are more than 100 conditions that are referred to as Arthritis. Arthritis is any combination of about 200 rheumatic diseases that affect the joints and the tissues that surround them. It causes inflammation and stiffness in the tissue around the joints.

Arthritis is as unavoidable as it is uncomfortable, often resulting in severe symptoms, like morning joint stiffness, persistent joint pain, carpal tunnel syndrome, tingling or numbness in extremities, plantar fasciitis, locked joints, and injuries that dont heal properly. The symptoms of Arthritis vary and sometimes come on gradually, as a group of traits. Those can include symptoms include swelling, inflammation, or stiffness anywhere on the body where joints meet connective tissue, especially upon awakening.

Arthritis can come on with any level of severity at any stage of life, but it most commonly affects aging citizens. Genetics is a prime cause of Arthritis, but it can also come from joint damage, obesity, infections, and repetitive motion over the years. Many cases are a combination of factors. In the United States, Arthritis is a primary cause of disability, leaving many unable to work for a living or even perform their household tasks.

The list below tells the common causes of 7 principal Arthritis categories:

- Inflammatory Arthritis is a cluster of diseases that include Psoriatic Arthritis, Rheumatoid Arthritis, Juvenile Idiopathic Arthritis, Systemic Lupus Erythematosus (lupus), and Ankylosing Spondylitis
- Degenerative or mechanical Arthritis, like Osteo-Arthritis, is caused by obesity, although it could be the result of congenitally abnormal joints.
- Soft tissue musculo-skeletal pain is from overuse, such as ”tennis elbow.”
- Back pain can be caused by any number of things, from a car accident to overuse.
- There are more than 200 disorders associated with connective tissue disease and the causes vary by the different types
- Infectious Arthritis occurs when the joints are affected by infection by bacteria, viruses (STDs, Hepatitis C) or fungi and parasites
- Metabolic Arthritis is caused by excess uric acid, as in the case of gout. Pseudogout occurs when calcium pyrophosphate (CPP) forms crystals in the cartilage of the wrists, knees, hips, shoulders, elbows, finger joints, toes, or ankles.

In the sections to come, the paper has been structured as follows. Section 2 gives an introduction to the data source from which the data has been acquired from for the analysis, and the highlights of the analysis performed. Section 3 describes the Descriptive Analysis performed, Section 4 describes the Statistical Analytics done, Section 5 describes the Machine Learning models run on the dataset, and finally, the last section comprises of the Conclusion.

2 Data Source and Analyses Performed

2.1 Data Source and Description

A crowdsourced dataset extracted from industry leading community sites was used as the primary data source for our analysis. It provides evidence based features from more than 190,000 product reviews for over 3,000 phenotypes.

Specifically, the full dataset includes:

- 193,000 Total User Submitted Reviews
- 4,201 Unique Cannabis Strains
- 19 Direct Psychological and Physiological Effects
- 44 Medical Related Conditions
- 6 Major Psychological and Physiological side effects
- CBx Cannabinoids Dataset:

- CBD
- CBDa
- CBC
- CBG
- CBGa
- CBDv
- CBDva

2.2 Analyses Performed

- Statistical analysis to determine how CBD percentages in the strains drive the efficacy of strains for one specific medical condition. Statistical analysis is highly relevant to determine how independent and dependent variables affect the user scoring and to define the customer sentiment sensitivity analysis.
- Natural language processing (NLP) was used to extract from the text reviews specific medical conditions, side effects, efficacy and sentiment. This information was then aggregated and used for training the machine learning models which are used to predict the ideal compound combinations for consumers with Arthritis.
- We used a variety of machine learning models to determine which strains have the highest efficacy with the least side effects for Arthritis. Model evaluation was based on model accuracy, outlier analysis, sample density/sparsity and generally accepted principles of algorithmic modeling.
- The dataset includes CBD, CBN, CBC, and CBG data. We did not include any terpene information since it was not available in this data extract. We have created a replicable process where information can easily be fed into our analysis engine and new insights can be derived quickly when terpene information becomes available.

3 Descriptive Analytics

3.1 Data Visualization

Our initial step was to present the current data with the lens of Ratings and Votes. As with most online directories and e-commerce sites, Ratings and Votes provides us with the first insight on trends and overall usage of the strains

Table 1 summarizes the number of reviews and strains used in our Arthritis specific analysis.

TABLE 1
Table showing subset of data of strains used solely for Arthritis

Medical Condition	Total Reviews in Dataset	Unique strains used by customers for Arthritis
Arthritis	125, 714	282

Figure 1 below represents the top 10 strains used by consumers and relative side effect scores. Side effect score is imputed as a function of the types of side effects of the strains that customers are taking.

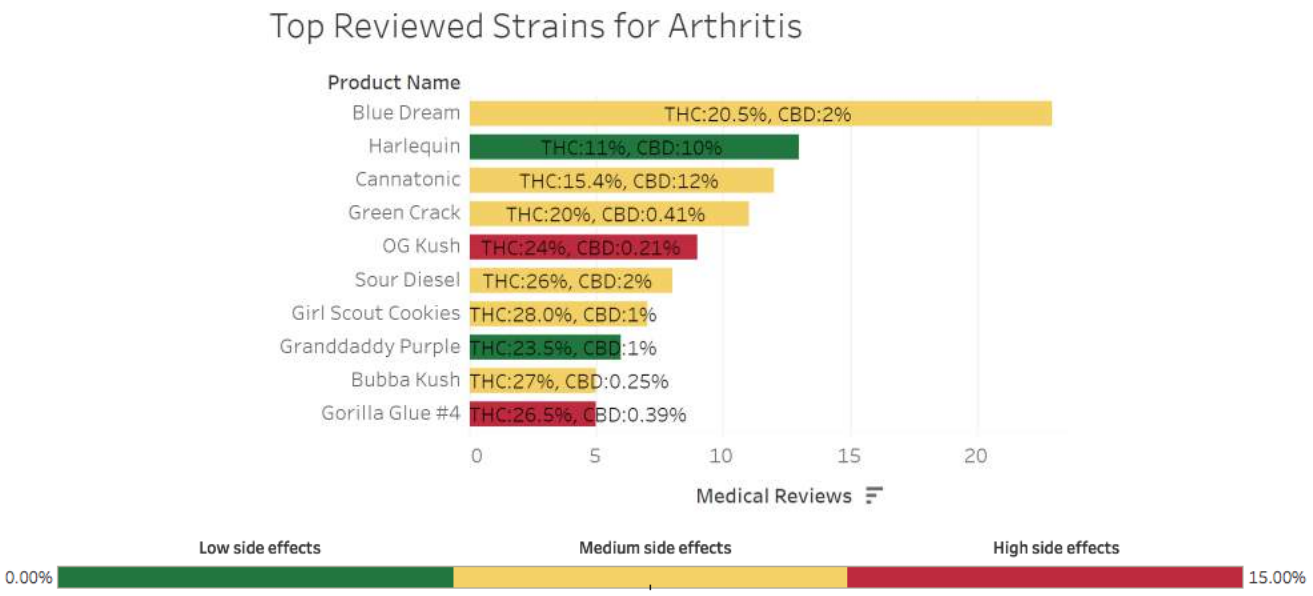

Fig. 1. Top Strains for Arthritis based on reviews and lowest mentioned side effects

3.2 Text Analytics

With more than 190,000 text reviews, it was necessary to identify the context within the text so that we can incorporate these as features in our machine learning models. The core dataset did not specifically identify Arthritis as a condition. However, consumers did mention in the reviews that they are either currently taking it for Arthritis or have taken it in the past with varying levels of effectiveness. The first step of our text analytics was developing a medical terminology corpus and using NLP (Natural Language Processing) engines to extract reviews with medical conditions only and tag these reviews with a binary value so that we can run statistical analysis on them. Assumption - Consumer reviews sometimes might be describing Arthritis as a medical condition, but do not directly have the word Arthritis in it. Such reviews are not being considered for the analysis.

*3.2.1 **Text Analysis***

We implemented a four step approach to text analysis. Each step and its rationale is described in detail below.

- **Step 1: Medical Lingua Corpus Creation**
  Consumer reviews from websites are intuitive for readers, but a system-based analysis to detect the right medical keywords and context requires a word / phrase dictionary. This corpus of medical marijuana / cannabis terms can be used as a filter for the reviews that have medical value. Furthermore, we can fine-tune the corpus to extract not just words, but phrases. For example, feeling paranoid and did not have paranoia like the last time reflect two opposite sentiments which if we used a standard keyword search algorithm would fail. Hence, contextual analysis is important for our next step of meta-tagging. The corpus data was extracted from these websites:
  - www.emedicinehealth.com
  - www.macmillandictionary.com
  - www.coloradopotguide.com

  Reviews that contain medical terms outside of the above listed corpora are not being captured.
- **Step 2: Medical Condition Matching**
  This step was to identify the medical terms present in each individual review addressing Arthritis. The count of each identified medical term is aggregated across all reviews. Based on the count, the identified medical terms are ranked to obtain the medical conditions that are strongly correlated with Arthritis.
- **Step 3: Sentiment Analysis and Keyword Extraction**
  Sentiment analysis was performed on the user reviews using the natural language toolkit and textblob package in Python. Based on the sentiment score derived from each review, reviews were classified into three groups - positive, negative and neutral. Emotions were derived in each review through an API call to IBM Watsons[4] natural language understanding. A combination of sentiment score, emotions and user ratings proved to be a better measure of efficacy of a strain in treating different medical conditions.
  Keywords were extracted from the reviews using rapid automatic keyword extraction algorithm using NLTK[5]. The results were visualized as a word cloud in Tableau to understand the frequently used words in reviews.

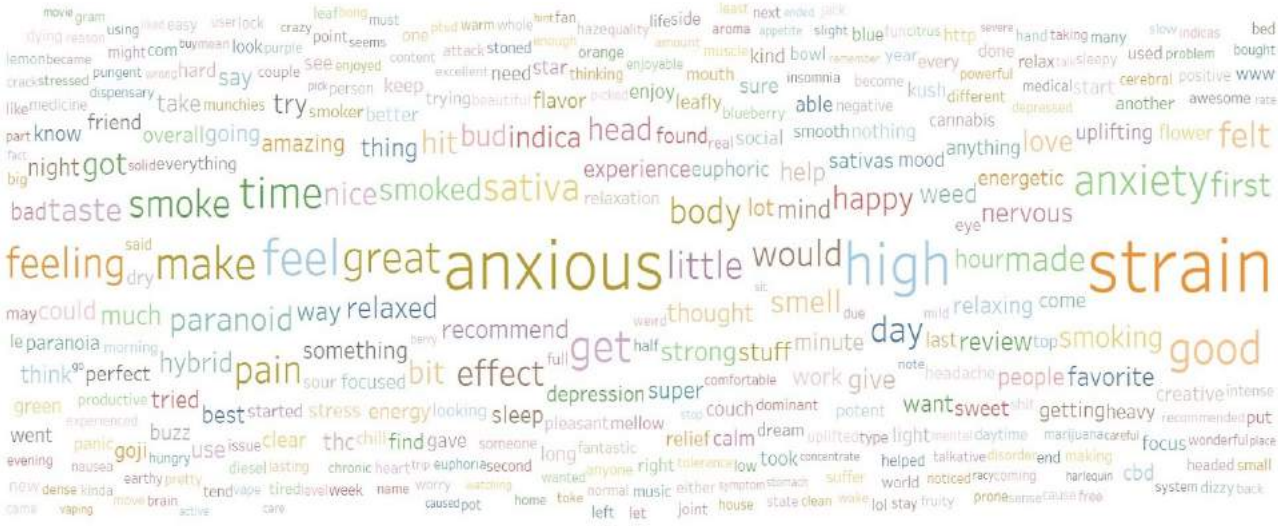

Fig. 2. Key Medical Conditions extracted based on magnitude and severity

- **Step 4: Visualization**

In this step, we wanted to visualize keywords and phrases in a network node graph so that we can understand the phrase traversal and direction aggregated from the reviews. Gephi[6] was used to visualize the results.

3.2.2 *Network Graph Visualization*

Figure 3 shows the Network Graph obtained.

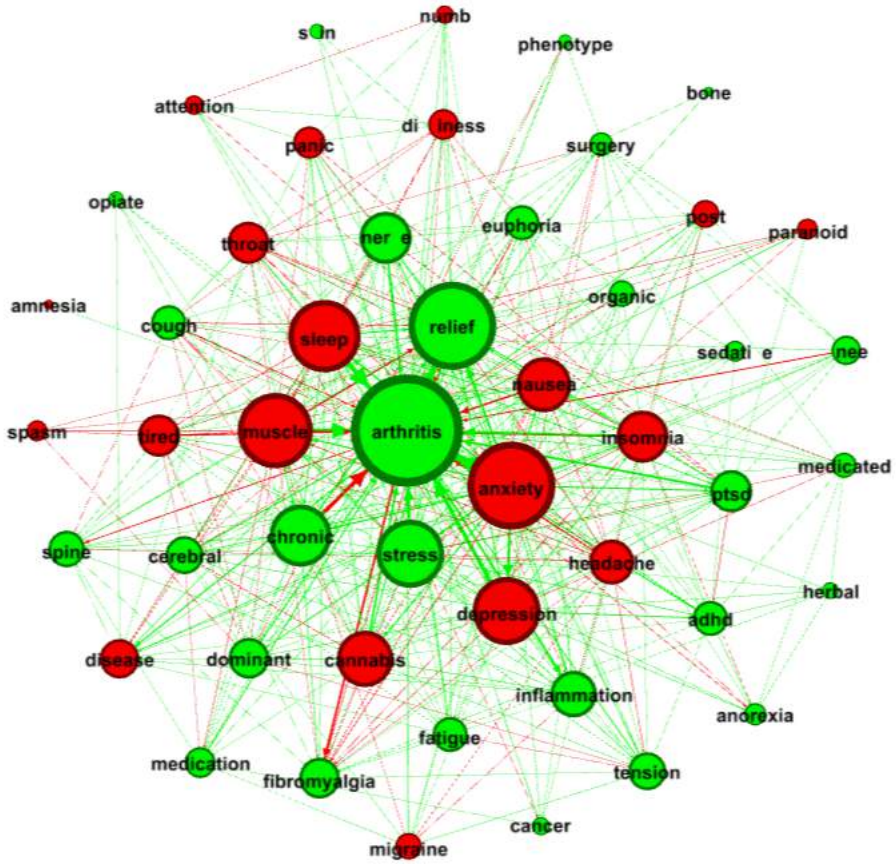

Fig. 3. Network Graph

- Nodes represent the medical terms that are present in the reviews related to Arthritis
- Edges/links connecting two medical terms represent the coexistence of the two medical terms in a single review
- Node size represents the frequency of that term appearing in the reviews
- Edge thickness is proportional to the count of co-occurrence of the two terms it connects
- Color represents average user ratings. Every review has a user rating associated with it. Mean value of ratings is 4.2. If a node is marked as green, the average ratings of reviews relating to that node is above mean. If the average rating is below mean, the node is highlighted in red

3.2.2.1 **Observations**: We observed that whenever a customer is talking about Arthritis in the reviews, there is strong correlation with other medical terms such as anxiety, depression, stress, insomnia and nausea.

3.2.3 *Dendrogram*

Results of the text analysis are visualized in the form of a dendrogram word tree to observe directionality. Font size of the text is proportional to the count (number of reviews which contain that particular medical term).

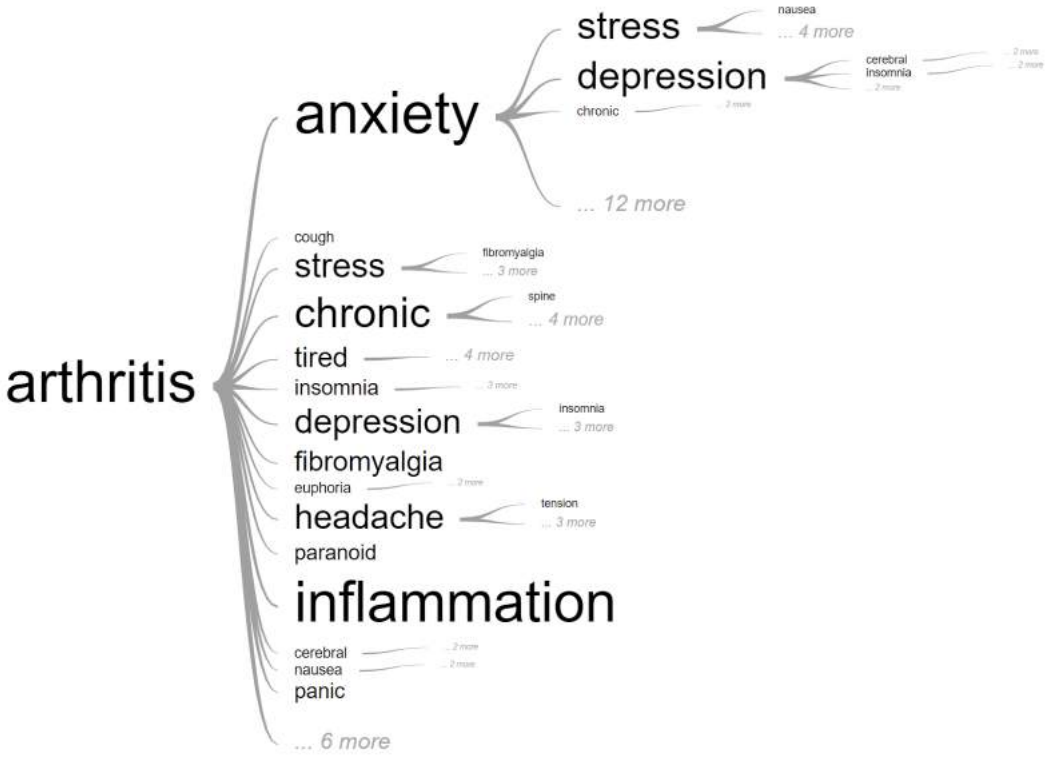

Fig. 4. Dendrogram of associated medical conditions

3.2.3.1 **Dendrogram Observation**: The analysis of the word cluster is as follows:

- Patients suffering from Arthritis are more likely to explicitly talk about inflammation, stress, depression, chronic pain, and fibromyalgia. While this may seem straightforward, the reverse may not be true. If we had started the node analysis with anxiety and depression, we may see Arthritis as a corollary condition, i.e. severity of the anxiety and depression is not effectively communicated by the consumers in their reviews
- When a patient has Arthritis and anxiety, they are more likely to talk about depression with insomnia and ADHD

4 Statistical Analysis

4.1 Data Preparation

This analysis examines how an increase or decrease in the percentage of non-psychoactive cannabinoids changes the number of arthritis-specific reviews. Non-psychoactive cannabinoids include CBD, CBDa, CBDv, CBDva, CBC, CBG, CBGa

CBD Total percentage is the sum of CBD and CBDa. CBG Total percentage is the sum of CBG and CBGa.

One dummy variable "D_Arthritis" is based on the number of Arthritis reviews.

The dataset has 3,739 observations, and 14 variables.

TABLE 2
Data Description

Variables	Description
X4_arthritis	Number of reviews for Arthritis for each strain
D-Arthritis	Dummy variable of Arthritis (1 =users do mention Arthritis in the product review, 0 = no mention of Arthritis in reviews)
sativa	The strain is Sativa
indica	The strain is Indica
CBD	The approximate CBD percentage contained in the strain
CBDa	The approximate CBDa percentage contained in the strain
CBDv	The approximate CBDv percentage contained in the strain
CBDva	The approximate CBDva percentage contained in the strain
CBC	The approximate CBC percentage contained in the strain
CBG	The approximate CBG percentage contained in the strain
CBGa	The approximate CBGa percentage contained in the strain
CBD_total	The total CBD percentage contained in the strain. Calculated using formula: $CBD_{total} = 0.877 * CBD_a + CBD$
CBG_total	The total CBG percentage contained in the strain. Calculated using formula: $CBG_{total} = 0.877 * CBG_a + CBG$

TABLE 3
Statistical Summary of Data

Statistic	N	Mean	Median	Min	Max
X4_arthritis	3739	2.27	1.0	0.00	23.00
D-Arthritis	3739	0.53	1.0	0.00	1.00
sativa	3739	0.45	0.5	0.00	1.00
indica	3739	0.55	0.5	0.00	1.00
CBD	3739	0.0001581	0.00001	0.00	0.0541
CBDa	3739	0.002688	0.0006	0.00	0.1836
CBDv	3739	7.057e-6	1e-6	1e-6	1.9e-3
CBDva	3739	5.876e-5	0.0	0.0	1.190e-2
CBC	3739	0.000145	0.0000	0.00	0.0042
CBG	3739	0.001029	0.0009	0.00001	0.0226
CBGa	3739	0.007	0.0057	0.00001	0.0466

4.2 Empirical Result and Discussion

4.2.0.1 **Correlation**: The first step is to observe the correlation between Arthritis and these cannabinoids, if there is some correlation, either positive or negative, then move on to the regression analysis as follows

4.2.1 Correlation Matrix

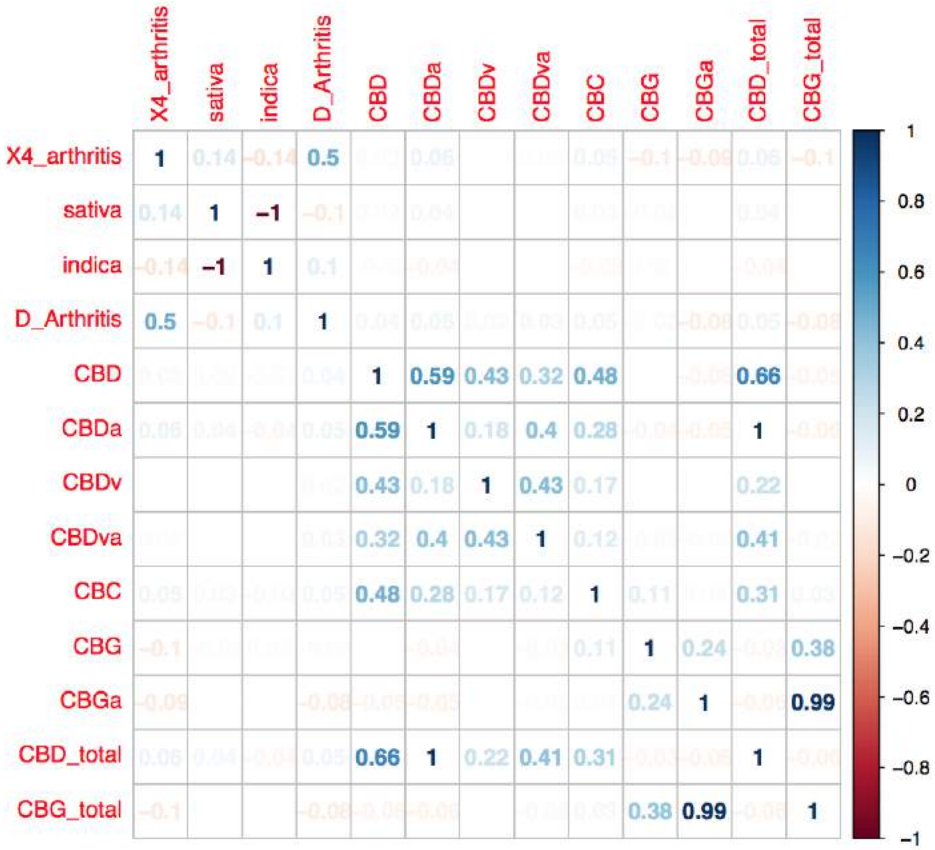

Fig. 5. Correlation Matrix

From Figure 5, we can see Arthritis number of reviews is positively correlated with CBDa, CBD_total, and CBC.

X4_Arthritis and CBDa: +0.06 X4_Arthritis and CBD_total: +0.06 X4_Arthritis and CBC: +0.05 Arthritis number of reviews is negatively correlated with CBG, CBGa and CBG_total X4_Arthritis and CBG: -0.1 X4_Arthritis and CBGa: -0.09 X4_Arthritis and CBG_total: -0.1

4.2.2 Non-Linear Regression Model: Arthritis vs. CBD_total and CBD_total^2

$$X4_{arthritis} = \beta_0 + \beta_1 \times CBD_{total} + \beta_2 \times CBD_{total}^2 + \epsilon \quad (1)$$

Here, β_0 is the intercept, β_1 , β_2 are the coefficient for each independent variable. ϵ is the error term, which includes all other factors that can affect the probability of consumers mentioning Arthritis in the product reviews, such as demographic variables. For example, consumer's age, BMI, gender.

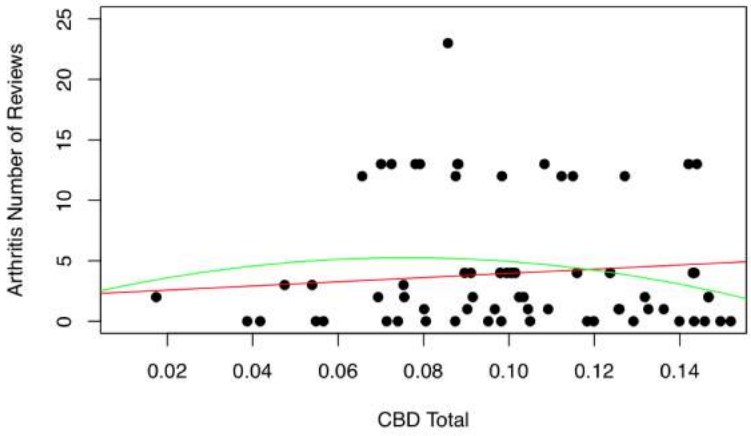

Fig. 6. Arthritis Number of Reviews Vs. CBD Total

The red-line is a prediction of change in Arthritis reviews according to change in CBD total percentage. It appears as an upward slope, indicating that as CBD total percentage increases, Arthritis number of reviews will increase. The green-line is a prediction of change in Arthritis reviews according to change in CBD total square. It appears there is a diminishing return, such that as CBD total reaches 0.075%, the continuing increase in CBD total will have a negative effect on Arthritis reviews.

Dependant Variable	$X4_{arthritis}$
CBD_{total}	80.98 (p=0.000204)***
CBD^2_{total}	-533.15 (p=0.002798)**
Intercept	2.19 (p<2e-16)***
Observations	3739

The regression results indicate both CBD total and CBD total square are significant.

4.2.3 *Non-Linear Regression Model: Arthritis vs. CBG_total and CBG_total^2*

$$X4_{arthritis} = \beta_0 + \beta_1 \times CBG_{total} + \beta_2 \times CBG^2_{total} + \epsilon \tag{2}$$

Here, β_0 is the intercept, β_1 , β_2 are the coefficient for each independent variable. ϵ is the error term, which includes all other factors that can affect the probability of consumers mentioning Arthritis in the product reviews, such as demographic variables. For example, consumer's age, BMI, gender.

Since most of the observations had Arthritis reviews = 0, which may skew the data, thus in this part of the analysis, X4_Arthritis=0 has been eliminated. After eliminating, this dataset contains 1,987 observations.

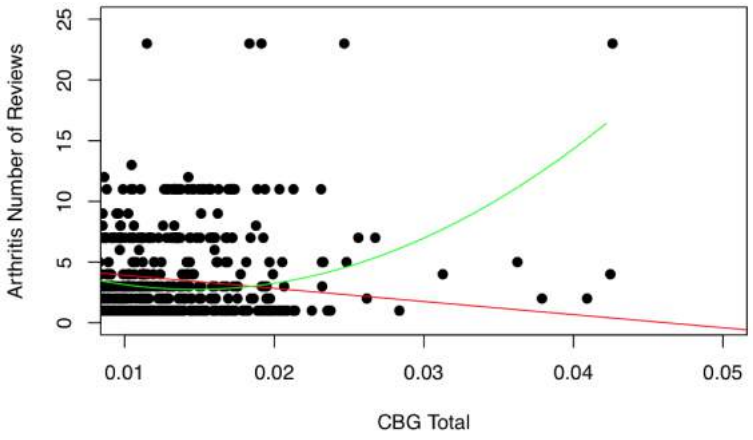

Fig. 7. Arthritis Number of Reviews Vs. CBG Total

The red-line is a prediction of change in Arthritis reviews according to change in CBG total percentage. It appears a downward slope, indicating that as CBG total percentage increases, Arthritis number of reviews will decrease. The green-line is a prediction of change in Arthritis reviews according to change in CBG total square. It appears there is a U-shape curve, as CBG total reaches 0.02%, the continuing increase in CBG total will have a positive effect on Arthritis reviews.

Dependant Variable	$X4_{arthritis}$
CBG_{total}	-532.61 (p<2e-16)***
CBG^2_{total}	18107.06 (p<2e-16)***
Intercept	6.67(p<2e-16)***
Observations	1987

The regression result indicates both CBG total and CBG total square are significant.

4.2.4 *Non-Linear Regression Model: Arthritis vs. CBC and CBC^2*

$$X4_{arthritis} = \beta_0 + \beta_1 \times CBCl + \beta_2 \times CBC^2 + \epsilon \tag{3}$$

Here, β_0 is the intercept, β_1 , β_2 are the coefficient for each independent variable. ϵ is the error term, which includes all other factors that can affect the probability of consumers mentioning Arthritis in the product reviews, such as demographic variables. For example, consumer's age, BMI, gender.

Since most of the observations has Arthritis reviews = 0, which may skew the data, thus in this part of the analysis, X4_Arthritis=0 has been eliminated. After eliminating, this dataset contains 1,987 observations.

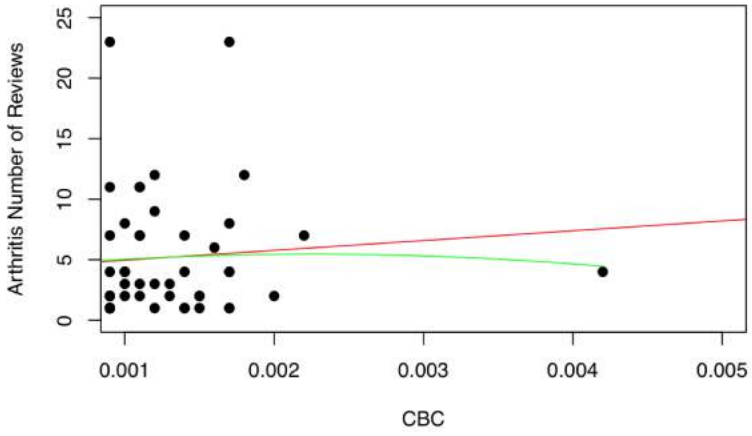

Fig. 8. Arthritis Number of Reviews Vs. CBC

The red-line is a prediction of change in Arthritis reviews according to change in CBC percentage. It also appears an upward slope, indicating that as CBC percentage increases, it has a positive effect on Arthritis number of reviews. The green-line is a prediction of change in Arthritis reviews according to change in CBC square. It shows there is a diminishing return.

Dependant Variable	$X4_{arthritis}$
CBC	1.193e+03 (p=0.0742)
CBC^2	-2.643e+05 (p=0.4597)
Intercept	4.110e+00(p<2e-16)***
Observations	1987

The regression summary shows only CBC percentage is significantly correlated with Arthritis number of reviews.

4.3 Statistical Analysis Results

- CBD Total (CBD & CBDa) - The probability that Arthritis appears in the customer review increases when CBD % increases. The optimal CBD total value for Arthritis from the reviews is observed to be **0.08%**.
- CBC - The probability that Arthritis appears in the customer review increases when when CBC % increases. CBC has a positive correlation with Arthritis number of reviews within 90% confidence interval, however the square term of CBC is not significantly correlated with Arthritis number of reviews.
- CBG - The probability that Arthritis appears in the customer review decreases as CBG % increases. The minimum amount of CBG total (a combination of CBG and CBGa) needed to use while treating Arthritis is observed to be 0.015%.

TABLE 4
Statistical Analysis Results

Cannabinoid	Unit increase in %	Optimum Value
CBD	Increase in number reviews	0.08%
CBC	Increase in number reviews	Not Significant
CBG	Decrease in number of reviews from up to 0.015%. Number of reviews increases from 0.015%	Minimum CBG value for positive correlation = 0.015%.

5 Machine Learning

5.1 Unsupervised Learning: Clustering Models

We used the following models to cluster the data to build a regression model more effectively. The lack of any training data means that we used a randomized training set from the core data. The details for each model are presented here:

- K-Means [7]
- Mini batch K-Means

5.1.1 Approach Summary

5.1.1.1 **Goal**: To see how CBD, weighted side effects score, efficacy score can be used to find patterns and cluster similar. A primal way of doing this is through an unsupervised learning approach that we explored using K-means. K-means is one of the most efficient algorithms in terms of performance and also gives a very easy understanding of what the data looks like.

5.1.1.2 **Assumptions**: The variance of distribution of each variable is spherical and also all variables have same variance.

- One of the important parameters in K-means is finding the K.
- The K can be found by elbow method through which K was found out to be 4.
- Weakness: Stability and choosing the value of K.

5.1.1.3 **Approach**: Features used for clustering - strain compositions (CBD %, THC %), negative side effects, number of reviews and weighed review ranking.

Step 1: Handling missing values - Since the missing values are less than 3% of the total records, we chose to drop the missing records and not apply any imputation method. Step 2: Scaling input features - The negative side-effect columns are on a scale of 1 to 10 while the chemical compositions are on a scale of 0 to 1. To establish a uniform influence of all input features on the clustering technique, we are choosing to scale all features onto a single logarithmic scale. Step 3: Clustering

5.1.2 Results

Comparing the different models, mini batch K-means with 4 clusters has high values for Silhouette coefficient and low inertia. This model has the highest clustering accuracy and is considered for the final conclusion.

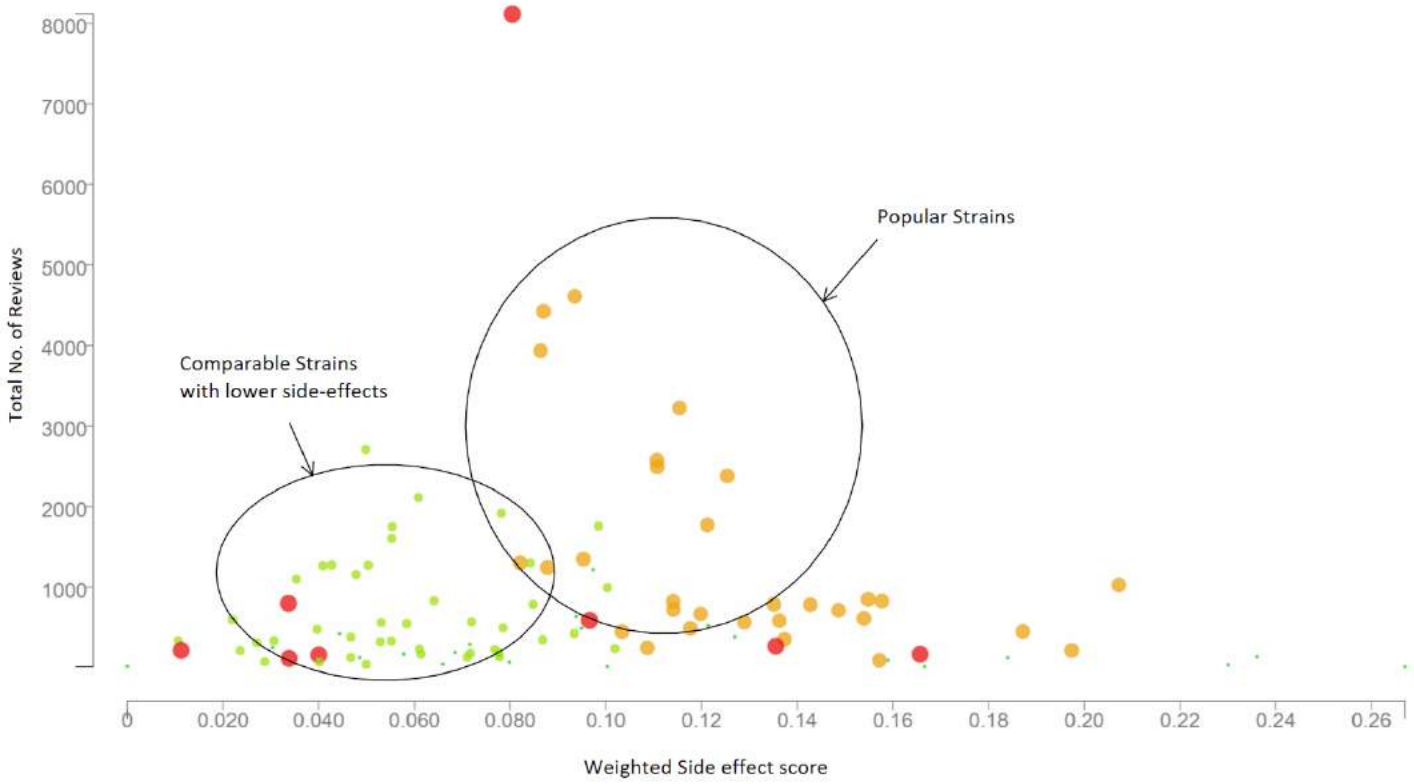

Fig. 9. Arthritis Number of Reviews Vs. Weighted Side-effect score

Figure 9 is a visualization of strains plotted against number of reviews along the y axis and weighted side-effects score along the x axis. Clusters are differentiated by coloring. Cluster highlighted in green is comprised of strains that can be used effectively for Arthritis with very low side effects. This cluster constitutes comparable strains that can be used as a substitute for the popular strains, with lower side-effects.

The top 10 comparable strains with low side-effect score and high reviews are shown in 5

TABLE 5
Comparable Strains

Sl. No	Strains	Total Reviews	Weighted Side-effect score
1	Granddaddy Purple	2708	0.05
2	Bubba Kush	2115	0.061
3	Northern Lights	1752	0.055
4	Blue Cheese	1606	0.055
5	Blackberry Kush	1275	0.043
6	Grape Ape	1267	0.041
7	Chemdawg	1270	0.05
8	Master Kush	1157	0.048
9	Cherry Pie	1105	0.035
10	Thin Mint Girl Scouts	447	0.04

5.2 Supervised Learning: Prediction and classification approach

Supervised Classification Techniques used:

- Artificial Neural Networks (ANN)
- Random Forests
- XGBoost
- Logistic Regression

5.2.0.1 **Goal**: Build a model that learns on the cannabinoid profile of cannabis strains and predict if a profile could be used to treat Arthritis. This model could be used to score any new strain and based on its chemical composition, predict if it can be used for Arthritis, or not.

5.2.0.2 **Approach Summaries**: The data used contains the complete cannabinoid profile of different strains along with the number of reviews for the strains and also information regarding whether Arthritis is mentioned as a medical condition in the reviews. If mentioned, the count of the reviews with Arthritis is recorded.

The target is a binary variable having a value of 1 if Arthritis is mentioned in the reviews for a given strain and 0 if Arthritis is not mentioned in any of the reviews. The input features are purely the cannabinoid profile of strains.

Data was randomly split into 70% training and 30% test. All the input features were transformed using standardization technique. Principal component analysis was used as the feature reduction technique. Various models were trained on cannabinoid profiles to predict the target binary variable (strain could be used for Arthritis or not).

5.2.1 Artificial Neural Networks

Neural Networks are a class of parametric models which are inspired by the functioning of the neurons. They consist of several "hidden" layers of neurons, which receive inputs and transmit them to the next layer, mixing the inputs and applying non-linearities, allowing for a complex decision function.

5.2.1.1 **Model Parameters**:

- Hidden Layers: 10
- Activation function: ReLU for hidden layers, Sigmoid for output layer
- Epoch: 5000
- Optimization Algorithm: Adam

5.2.1.2 **Results**: Model accuracy obtained is 68% and the confusion matrix is as shown below in 6

TABLE 6
Confusion Matrix for Neural Network

	Predicted (1)	Predicted (0)
Total	68	115
Ground Truth (1)	29	20
Ground Truth (0)	39	95

5.2.2 Random Forests

Decision tree classification is a simple algorithm which builds a decision tree. Each node of the decision tree includes a condition on one of the input features.

A Random Forest classifier is made of many decision trees. When predicting a new record, it is predicted by each tree, and each tree "votes" for the final answer of the forest.The forest chooses the class having the most votes.

When "growing" (i.e., training) the forest,

- For each tree, a random sample of the training set is used;
- For each decision point in the tree, a random subset of the input features is considered.

Random Forests generally provide good results, at the expense of "explainability" of the model.

5.2.2.1 **Model parameters**:

- Method: Bootstrap aggregation
- Variable importance metric: Gini
- Max tree depth: 15
- Min sample to split: 5

5.2.2.2 **Results**: Model accuracy obtained is 68% and the confusion matrix is as shown below in 7

TABLE 7
Confusion Matrix for Random Forests

	Predicted (1)	Predicted (0)
Total	71	112
Ground Truth (1)	31	18
Ground Truth (0)	40	94

5.2.3 XGBoost

XGBoost is an advanced gradient boosted tree algorithm. It has support for parallel processing, regularization, early stopping which makes it a very fast, scalable and accurate algorithm

5.2.3.1 **Model parameters**:

- Maximum number of trees: 300
- Maximum depth of tree: 3. (Maximum depth of each tree. High values can increase the quality of the prediction, but can lead to overfitting.)
- Learning rate: 0.2
- Subsample: 1 (Subsample ratio for the data to be used in each tree. Low values can prevent overfitting.)

5.2.3.2 **Results**: Model accuracy obtained is 69% and the confusion matrix is as shown below in 8

TABLE 8
Confusion Matrix for XGBoost

	Predicted (1)	Predicted (0)
Total	63	120
Ground Truth (1)	28	21
Ground Truth (0)	35	99

5.2.4 Logistic Regression

Despite its name, Logistic Regression is a classification algorithm, using a linear model (ie, it computes the target feature as a linear combination of input features). Logistic Regression minimizes a specific cost function (called logit or sigmoid function), which makes it appropriate for classification. A simple Logistic Regression algorithm is prone to overfitting and sensitive to errors in the input dataset. To address these issues, it is possible to use a penalty (or regularization term) to the weights.

5.2.4.1 **Model parameters**:

- 'L2' regularization is implemented.
- Penalty parameter C of the error term 0.1

5.2.4.2 **Results**: Model accuracy obtained is 57% and the confusion matrix is as shown below in 9

TABLE 9
Confusion Matrix for Logistic Regression

	Predicted (1)	Predicted (0)
Total	102	81
Ground Truth (1)	36	13
Ground Truth (0)	66	68

5.2.5 Model Comparison

Parameters used to compare models are:

- Precision - Proportion of positive predictions that were indeed positive (in the test set)
- Recall - Proportion of actual positive values found by the classifier
- F1 Score - Harmonic mean between Precision and Recall
- Accuracy - Proportion of correct predictions (positive and negative) in the test set
- ROC-AUC Score - Area under the ROC; from 0.5 (random model) to 1 (perfect model)
- Log loss - Error metric that takes into account the predicted probabilities (the lower the better)

TABLE 10
Model Comparison Table

Model	Precision	Recall	F1-Score	Accuracy	ROC-AUC Curve	Log Loss
Artificial Neural Networks	0.4265	0.5918	0.4957	0.6776	0.6924	0.5452
Random Forests	0.4366	0.6327	0.5167	0.6831	0.6697	0.5638
XGBoost	0.4444	0.5714	0.5000	0.6940	0.6777	0.6132
Logistic Regression	0.3529	0.7347	0.4768	0.5683	0.6692	0.5466

Random forest has the highest accuracy and F1 score. Given the data we are considering, this model does the best prediction and can be used for scoring new data.

The best accuracy that could be achieved is 68%. This is considering only the cannabinoid profile of the strain. We are expecting this to increase when we include terpene data in our analysis.

6 Conclusion

To summarize our findings:

Our text analysis of reviews reveals that whenever a customer mentions Arthritis in the reviews there is strong correlation with other medical terms such as anxiety, depression, stress and fibromyalgia. Specifically, stress and depression are the other top 2 keywords used in the Arthritis reviews. In addition, when they mention Arthritis and stress, they are more likely to talk about issues with insomnia, ADHD and nausea.

The Statistical Analysis shows the following:

- CBD Total (CBD and CBDa) - The probability that Arthritis appears in the customer review increases when CBD % increases. The optimal CBD total value for Arthritis from the reviews is observed to be 0.08%.
- CBC - The probability that Arthritis appears in the customer review increases when CBC % increases. CBC has a positive correlation with Arthritis number of reviews within 90% confidence interval, however the square term of CBC is not significantly correlated with Arthritis number of reviews.
- CBG - The probability that Arthritis appears in the customer review decreases as CBG % increases. The minimum amount of CBG total (a combination of CBG and CBGa) needed to use while treating Arthritis is observed to be 0.015%

Coming to the Machine Learning analysis, the following results were concluded:

- Random forest was observed to be the best model in predicting if a strain could be used for Arthritis or not. The obtained accuracy was 68% with an F-1 score of 0.52.
- Silhouette coefficient and inertia was used to compare different clustering approaches. Mini-batch k means with 4 clusters had the highest Silhouette coefficient with a value of 0.3027 and an inertia value of 167.2.
- The top 5 strains with high efficacy while treating Arthritis are Granddaddy Purple, Bubba Kush, Northern Lights, Blue Cheese, Blackberry Kush.

REFERENCES

[1] S. Kutty, C. Cheney, et.al *Evidence Based Approach Using Multivariate Regression Analysis And Machine Learning Algorithms To Discover Optimum Medical Cannabis Compound Ratios For Highest Efficacy And Lowest Side Effects In PTSD Subjects* (2017)

[2] Blessing, Esther M., Maria M. Steenkamp, Jorge Manzanares, and Charles R. Marmar. *"Cannabidiol as a Potential Treatment for Anxiety Disorders."* Neurotherapeutics 12.4 (2015): 825-36. Print

[3] Schier, Alexandre Rafael De Mello, Natalia Pinho De Oliveira Ribeiro, Adriana Cardoso De Oliveira E Silva, Jaime Eduardo Cecilio Hallak, Jos Alexandre S. Crippa, Antonio E. Nardi, and Antonio Waldo Zuardi. *"Cannabidiol, a Cannabis Sativa Constituent, as an Anxiolytic Drug."* Revista Brasileira De Psiquiatria 34 (2012). Print.

[4] Petrick, S. R. *"On Natural Language Based Computer Systems."* IBM Journal of Research and Development 20.4 (1976): 314-25. Print.

[5] Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. OReilly Media Inc.

[6] Bastian M., Heymann S., Jacomy M. (2009). *Gephi: an open source software for exploring and manipulating networks.*International AAAI Conference on Weblogs and Social Media

[7] "Jin, Xin ("2010"). *"K-Means Clustering"*. pp 563-564, 10.1007/978-0-387-30164-8_425,"Springer US"

[8] Wooldridge, J. M., & Wooldridge, J. M. (2003). *Solutions manual and supplementary materials for Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

[9] Cramer, J. S. (1994). *The LOGIT model: an introduction for economists.* London: Arnold.