

# Drug Reviews: Sentiment Analysis

Group 5: Ahmed Farid Khan, Ananya Anand, Aryan Kumar, Srithijaa Sankepally

Link to Repository: <https://github.com/aryan-bu/BA820>

Our dataset consists of the reviews and rating by the users for particular drugs and conditions. After performing the necessary cleaning, preprocessing and analysis, these are our findings.

## Exploratory Data Analysis (EDA)

We analyzed customer satisfactions based on product reviews. From Fig. 1, we can infer that there is a satisfaction among the raters, with more individuals or instances favoring higher ratings. The highest rating (10) is significantly more common than neutral ratings (4-6.9) and the bad ratings (1-3.9). The distribution is skewed towards higher ratings, indicating that the drug which is being rated is generally viewed positively by the customers or has a tendency to receive higher evaluation scores.

From Fig. 2, we can see that the data points show considerable variability, with spikes and drops in review counts throughout the years. This could indicate periods of heightened interest or activity due to specific events, marketing campaigns and product launches. A lot of the spikes were towards the end of the year which might be due to seasonal fluctuations. The highest concentration of reviews occurs between 2016 and 2017, suggesting a peak in activity or interest. This period might have seen the most significant engagement from users. After reaching a peak, there's a noticeable decline in the number of reviews, particularly sharp towards 2018. This decline could suggest changes in market conditions, or possibly issues with the product leading to lower usage or satisfaction.

The most common conditions that the reviewers talked about were 'Birth Control' and 'Depression' (Fig. 3). Elinest was the most highly rated drug for birth control (Fig. 4) with the worst one being Dasetta 7 / 7 / 7 (Fig. 5).

## Preliminary Analysis

For our sentiment analysis, we converted ratings data into sentiment data so that the following ratings are classified as the following sentiments:

1. (1- 3): Negative
2. (4-6): Neutral
3. (7-10): Positive

Once we have these labels, we use them in supervised classification models and use accuracy and confusion matrices to assess how good the model is performing.

One caveat that we address is the relative class imbalance in our dataset. We do not have a lot of ratings between 4 and 6 so we have relatively few reviews labeled as neutral reviews. As a consequence we used SMOTE to oversample the minority class.

Before running the model on our entire dataset, we sampled 2000 rows within this preliminary analysis to reduce the computational cost. We created embeddings using Bag-of-Words (BoW) and used these embeddings as the features in our classification models.

We ran three classification models (logistic regression, random forest, gradient boosting) with and without oversampling and the accuracy results that we obtained are displayed in the table below:

	No Minority Oversampling	SMOTE
<b>Logistic Regression</b>	69.75%	61.75%
<b>Random Forest</b>	67%	67%
<b>Gradient Boosting</b>	64.25%	61.25%

From our table, we can see that logistic regression without minority oversampling results in the highest accuracy of 69.75%. We can also see that the highest accuracy after doing SMOTE is produced by the random forest classification. However, as displayed in the appendix (Tables 1-6), it appears that Logistic Regression with SMOTE actually does the best job at not classifying most of the neutral and negative responses as positive (the majority class). So, without any hyperparameter tuning, it seems that logistic regression with SMOTE might be a good fit for this classification problem.

### Next Steps:

As part of our final analysis, we will conduct the following to improve the accuracy of our model:

1. We will lemmatize the text data to capture the meaning of the words. We will use lemmatization rather than stemming because we want to capture the meanings of the words.
2. We will try other methods of embedding. We will try both Word2Vec and n-grams. We believe Word2Vec will perform better since we are more interested in the semantic meaning rather than the local context of words. However, Word2Vec is more computationally expensive so if we are not able to do Word2Vec we will consider n-grams.
3. We will tune the models so that they give us a better balanced accuracy compared to what we are able to achieve right now.
4. We will use the whole dataset rather than sampling it. So that we have more data for the model to be trained and tested on.

## Appendix

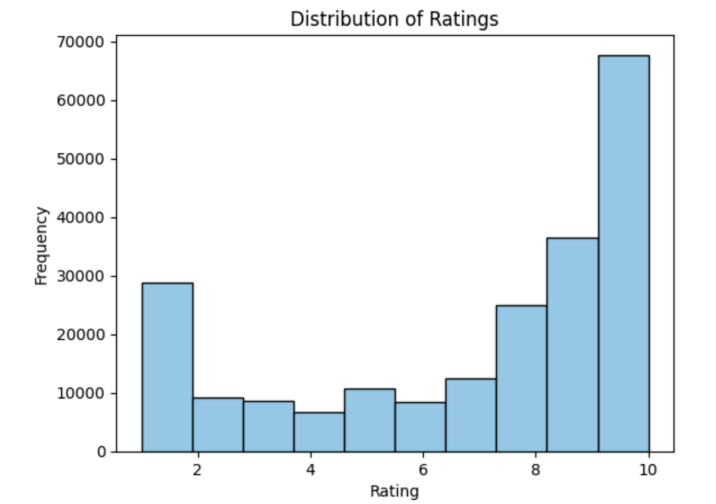


Fig. 1

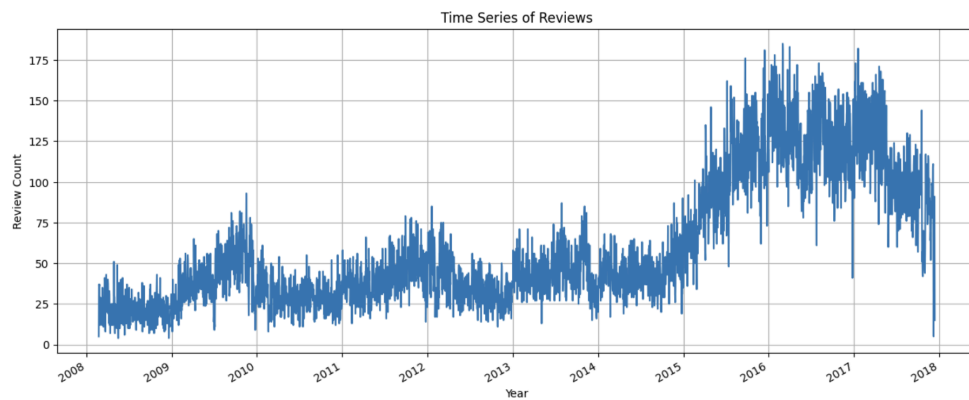


Fig. 2

Birth Control	38436
Depression	12164
Pain	8245
Anxiety	7812
Acne	7435
Bipolar Disorde	5604
Insomnia	4904
Weight Loss	4857
Obesity	4757
ADHD	4509

Name: condition, dtype: int64

Fig. 3

Worst 5 drugs for 'Birth Control':  
['Dasetta 7 / 7 / 7', 'Larin 24 Fe', 'Philith', 'Pirmella 1 / 35', 'Tri-Lo-Estarylla']

Worst 5 drugs for 'Depression':  
['Vyvanse', 'Luvox CR', 'Fetzima', 'Levomilnacipran', 'Pamelor']

Worst 5 drugs for 'Pain':  
['Capsin', 'Capzasin', 'Dolobid', 'Prialt', 'Ziconotide']

Worst 5 drugs for 'Anxiety':  
['Hyzine', 'Vanspar', 'Compazine', 'Prochlorperazine', 'Trileptal']

Worst 5 drugs for 'Acne':  
['Tri-Estarylla', 'Junel Fe 1.5 / 30', 'Loestrin Fe 1 / 20', 'Nortrel 1 / 35', 'Microgestin Fe 1.5 / 30']

Worst 5 drugs for 'Bipolar Disorder':  
['Carbatrol', 'Risperdal Consta', 'Celexa', 'Depakene', 'Equetro']

Worst 5 drugs for 'Insomnia':  
['ZzzQuil', 'Belsomra', 'Suvorexant', 'Valerian', 'Ramelteon']

Worst 5 drugs for 'Weight Loss':  
['Contrave', 'Bupropion / naltrexone', 'Fastin', 'Ionamin', 'Qsymia']

Worst 5 drugs for 'Obesity':  
['Methylphenidate', 'Alli', 'Bontril PDM', 'Chorionic gonadotropin (hcg)', 'Contrave']

Worst 5 drugs for 'ADHD':  
['Tenex', 'Budeprion XL', 'Imipramine', 'Mydayis', 'Quillivant XR']

Fig. 4

Best for 'Birth Control':  
['Elinest', 'Low-Ogestrel-21', 'Plan B', 'Zovia 1 / 50', 'Heather']

Best for 'Depression':  
['Asendin', 'Forfivo XL', 'Invega', 'Lamictal DS', 'Niravam']

Best for 'Pain':  
['Anacin', 'Anaprox-DS', 'Aspirin / caffeine', 'Bupivacaine liposome', 'Buprenex']

Best for 'Anxiety':  
['Alprazolam Intensol', 'Aspirin / meprobamate', 'Diazepam Intensol', 'Lorazepam Intensol', 'Micrainin']

Best for 'Acne':  
['Acnex', 'Avar', 'Avita', 'BenzEfoam Ultra', 'Benzoyl peroxide / hydrocortisone']

Best for 'Bipolar Disorder':  
['Eskalith-CR', 'Klonopin Wafer', 'Tiagabine', 'Nuvigil', 'Armodafinil']

Best for 'Insomnia':  
['Ethchlorvynol', 'Hetlioz', 'Nembutal Sodium', 'Pentobarbital', 'Remeron SolTab']

Best for 'Weight Loss':  
['Megace', 'Suprenza', 'T-Diet', 'Megace ES', 'Phentercot']

Best for 'Obesity':  
['Belviq XR', 'Desoxyn', 'Fastin', 'Ionamin', 'Methamphetamine']

Best for 'ADHD':  
['Dextrostat', 'ProCentra', 'Selegiline', 'Desoxyn', 'Cylert']

Fig. 5

## LogReg without SMOTE

Accuracy: 0.6975			
	<b>negative</b>	<b>neutral</b>	<b>positive</b>
<b>negative</b>	0.520833	0.052083	0.427083
<b>neutral</b>	0.212766	0.063830	0.723404
<b>positive</b>	0.070039	0.050584	0.879377

Table 1

## LogReg with SMOTE

Accuracy with SMOTE: 0.6175			
	<b>negative</b>	<b>neutral</b>	<b>positive</b>
<b>negative</b>	0.572917	0.093750	0.333333
<b>neutral</b>	0.255319	0.170213	0.574468
<b>positive</b>	0.128405	0.155642	0.715953

Table 2

### Random Forest without SMOTE

Accuracy: 0.67			
	negative	neutral	positive
negative	0.187500	0.000000	0.812500
neutral	0.127660	0.000000	0.872340
positive	0.023346	0.003891	0.972763

Table 3

### Random Forest with SMOTE

Accuracy with SMOTE: 0.67			
	negative	neutral	positive
negative	0.447917	0.052083	0.500000
neutral	0.148936	0.106383	0.744681
positive	0.077821	0.066148	0.856031

Table 4

### Gradient Boost without SMOTE

Accuracy: 0.6475			
	negative	neutral	positive
negative	0.145833	0.041667	0.812500
neutral	0.106383	0.000000	0.893617
positive	0.038911	0.007782	0.953307










Table 5

### Gradient Boost with SMOTE

Accuracy with SMOTE: 0.6125			
	negative	neutral	positive
negative	0.437500	0.125000	0.437500
neutral	0.170213	0.276596	0.553191
positive	0.128405	0.132296	0.739300

Table 6

## Group Contributions

	Title	...	Assignees	...	Status	...
1	Combine + Sort + Sample #3		 aryan-bu	▼	Done	▼
2	Preprocessing - Missing Values + Cleaning #4		 aryan-bu	▼	Done	▼
3	Tokenization- sentence, word, punctuation #5		 ananyaand0501	▼	Done	▼
4	Sentiment Segmentation #6		 ananyaand0501	▼	Done	▼
5	EDA (Srithijaa) #7		 Srithijaa11	▼	Done	▼
6	Vectorization (Srithijaa) #10		 Srithijaa11	▼	Done	▼
7	Analysis (Ahmed) #8		 ahmedfaridkhan	▼	Done	▼
8	Classification Models #11		 ahmedfaridkhan	▼	Done	▼
9	Common Conditions (Srithijaa) #12		 Srithijaa11	▼	Done	▼