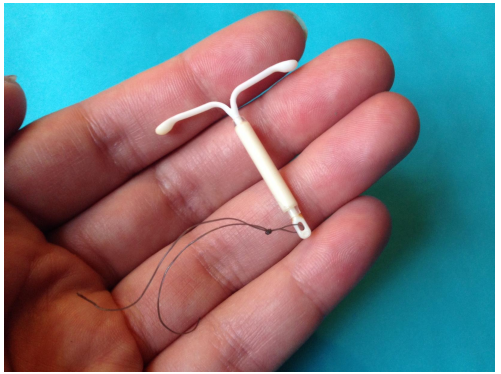
IOD_mini_project3

Ng Jing Kang

IOD_mini_project3 - Overview

Overview	
Background	Data analyst at a Medical institute/Pharmaceutical company
Purpose	POC on patient feedback analysis on the drugs (Improve on drug development/marketing)
Stakeholders	Audience



IOD_mini_project3 - Dataset

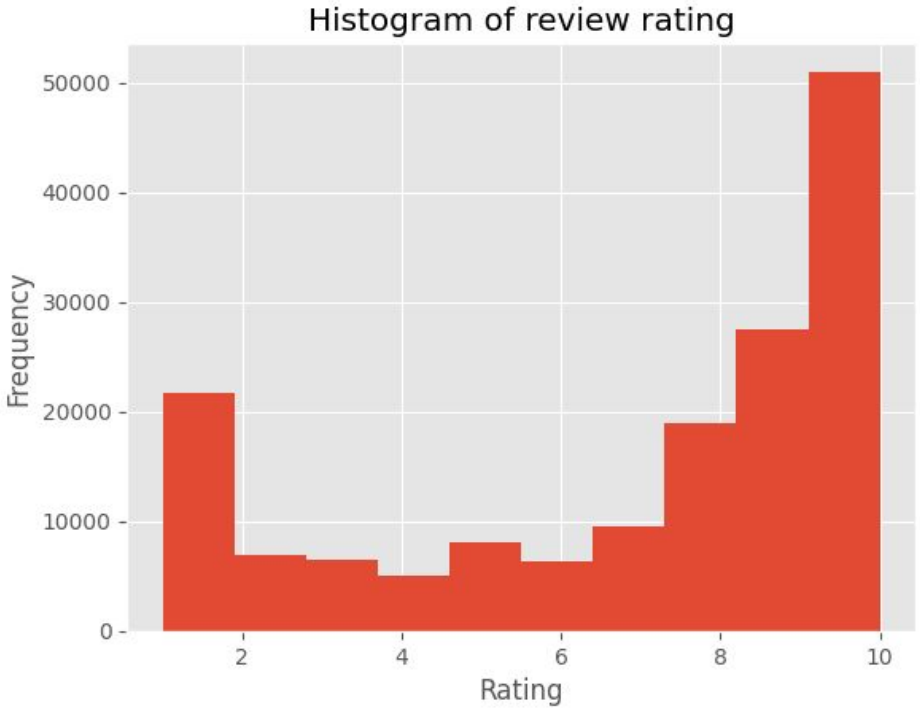
Source: https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon-winter-2018?select=drugsComTrain_raw.csv

Columns	Description	Example
uniqueID	Unique ID for review entry	206461
drugName	Name of drug used by patient	Valsartan
condition	Name of condition experienced by patient	Left Ventricular Dysfunction
review	Patient review (format: written text)	"It has no side effect, I take it in combination of Bystolic 5 Mg and Fish Oil"
rating	Rating score (1-10 stars)	9
date	Date of review entry	20-May-12
usefulCount	Count of users who found this review helpful	27

IOD_mini_project3 - Initial EDA

Columns	Example	EDA
uniqueID	206461	NA
drugName	Valsartan	3436 unique drugs
condition	Left Ventricular Dysfunction	885 unique condition, 1799 missing values
review	"It has no side effect, I take it in combination of Bystolic 5 Mg and Fish Oil"	NA for now. Will be useful during Sentiment analysis
rating	9	Histogram plot (next slide)
date	20-May-12	NA (Not focusing on time series)
usefulCount	27	NA for now. Will be useful during Sentiment analysis

IOD_mini_project3 - Initial EDA

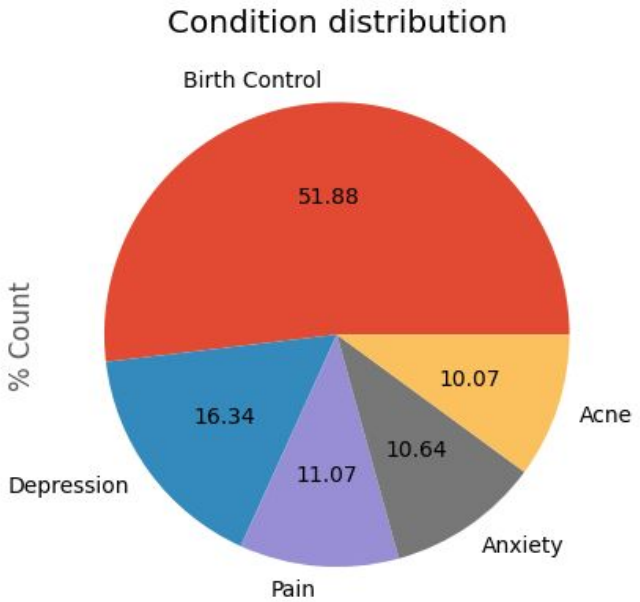


IOD_mini_project3 - Use Case (Classification)

Predict condition based on reviews:

- Application 1:
  - Recall 1799 missing values found in data
  - Condition input could be non-mandatory
  - Input can be disrupted due to technical fault on surveying platform
- Application 2:
  - After training model, predictions can be done on medical related reviews to predict conditions
  - This can aid diagnostic department

In relation to both applications, this model can be used in conjunction with speech to text models to generate textual review (e.g. patient is not tech savvy)



IOD_mini_project3 - Use Case (Classification)

Comparison of classifiers:

Classifier	Count Vectors	WordLevel TF-IDF	N-Gram Vectors	CharLevel Vectors	Mean Accuracy (3.s.f)
BernoulliNB	0.882061	0.871700	0.796018	0.785566	0.834
Logistic Regression	0.939814	0.927561	0.885215	0.921164	0.918
SVC	0.916389	0.950536	0.920623	0.942607	0.933
RandomForestClassifier	0.933508	0.939094	0.905937	0.933958	0.928

Findings and Improvements:

- Similar scores for different classifiers, however, SVC has the best mean accuracy
- Look into computational resource spending (NB used alot of memory, SVC took alot of time training)
- Use a more powerful model (NN)

IOD_mini_project3 - Sentiment Analysis

Sentiment Analysis:

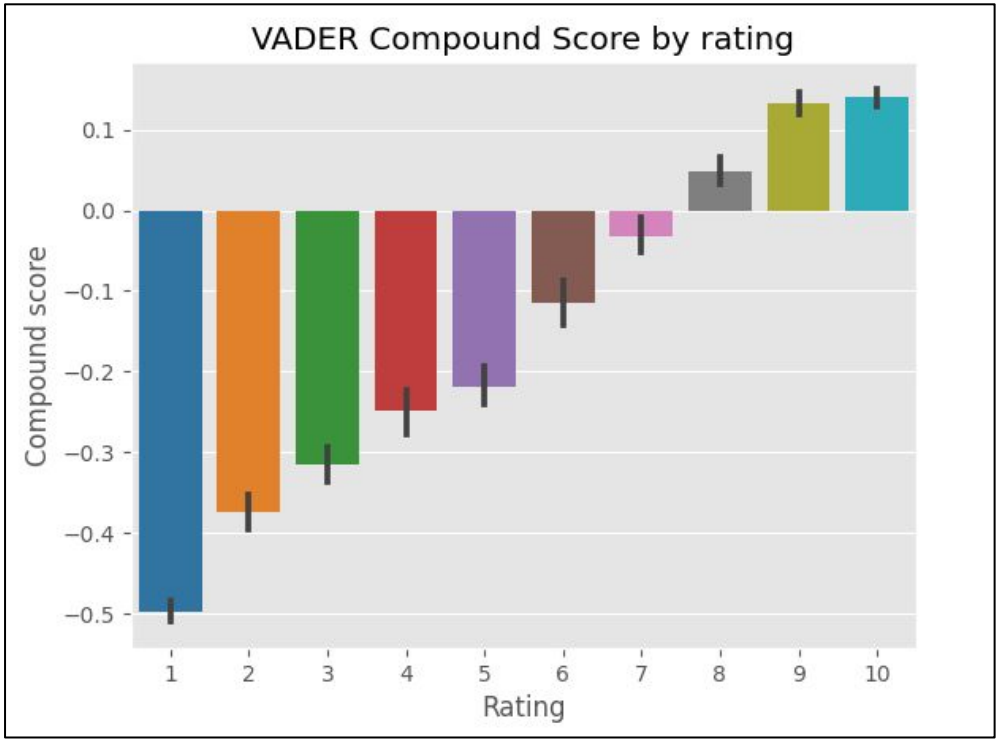
- Determine sentiment intensity
  - Determine, at a glance, which statements are positive/neutral/negative
  - Validate if scoring is accurate (1-10 stars) (Useful for marketing strategy, e.g. weed out “trolling” comments)
  - If no scoring is available, we can determine the sentiment of the review and choose to act on it (If largely positive - reinforce marketing around it, if largely negative - improve on the drug)

- Recall useful count column
  - Determine keywords that are helpful to users (Useful for marketing strategy, e.g. labelling of drug products/advertising)

IOD_mini_project3 - Sentiment Analysis

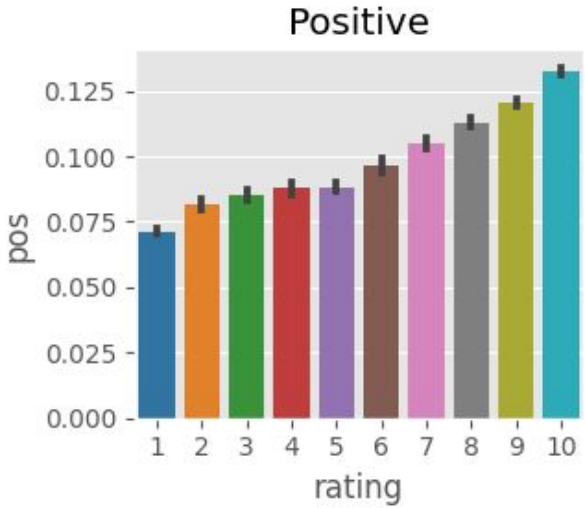
VADER (Valence Aware Dictionary for Sentiment Reasoning):

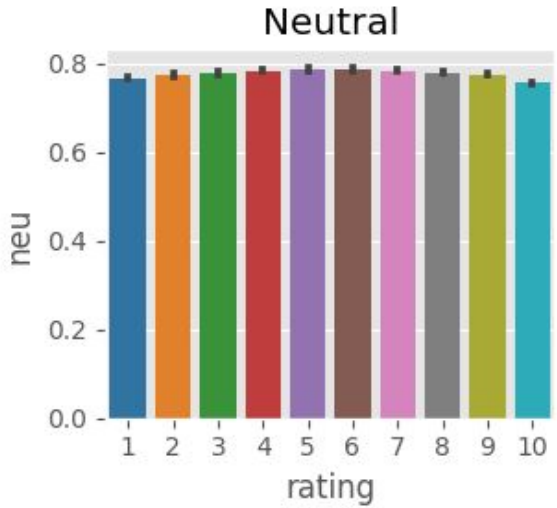
Sentiment intensity	Description
Compound score	Overall aggregated score
Pos	Positivity score
Neu	Neutrality score
Neg	Negativity score

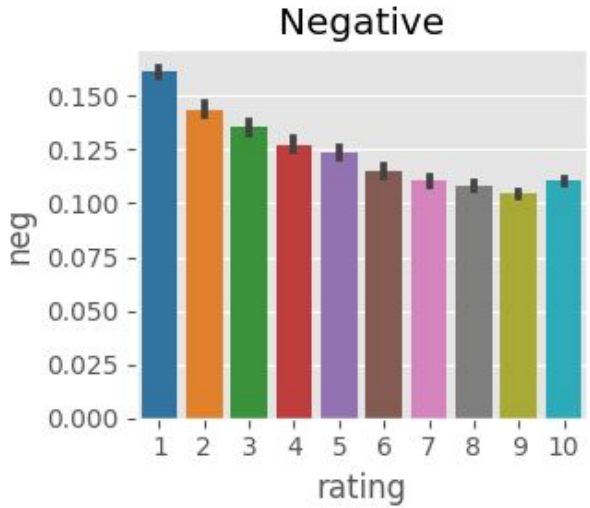


IOD_mini_project3 - Sentiment Analysis

VADER (Valence Aware Dictionary for Sentiment Reasoning):

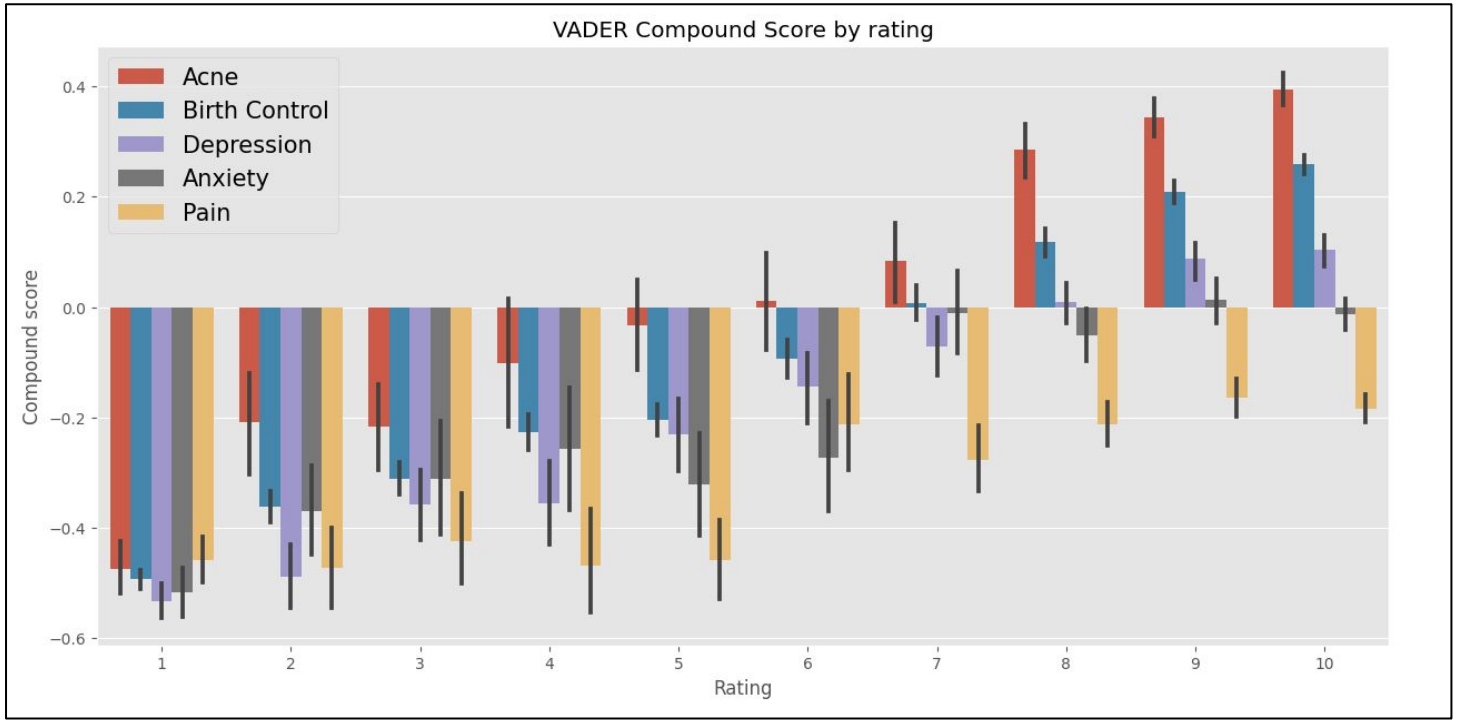






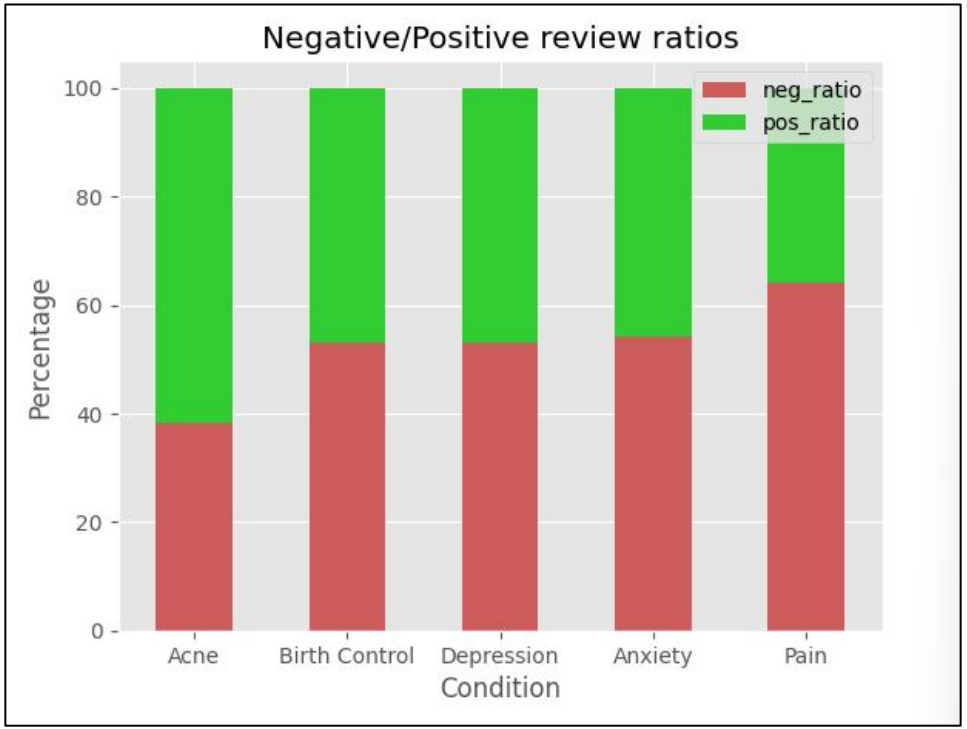
IOD_mini_project3 - Sentiment Analysis

VADER (Valence Aware Dictionary for Sentiment Reasoning):



IOD_mini_project3 - Sentiment Analysis

VADER (Valence Aware Dictionary for Sentiment Reasoning):



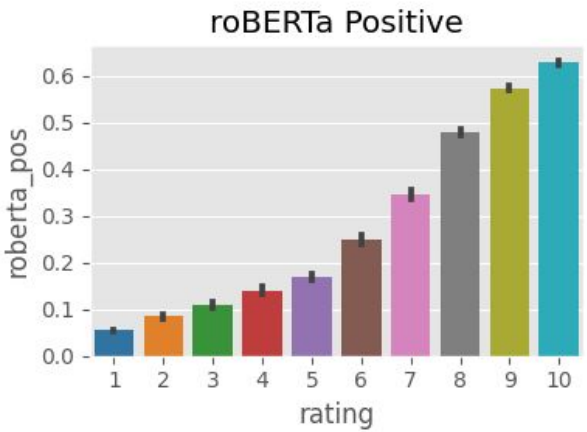
IOD_mini_project3 - Sentiment Analysis

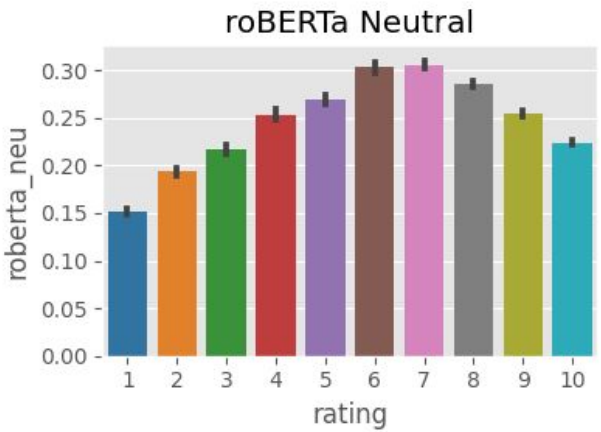
VADER vs Hugging face - roBERTa (Robustly Optimized BERT Pre-training Approach):

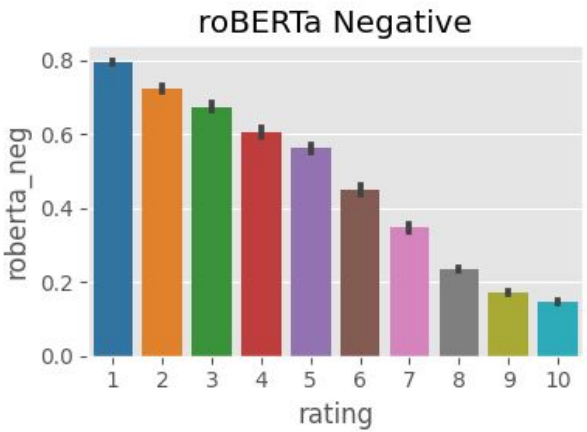
Example reviews	Rating	vader_compound	roBERTa_Pos	roBERTa_Neu	roBERTa_Neg
"I've tried numerous meds for back pain and neurogenic pain from MS, Opana works good, although the new formula made me very sick, thank God generic are not the plastic tablets. 3X20 per day and 1-3 Norco for BT. I use Fentanyl 25mcg on weekend instead of OP to give my digestive tract a break."	7	-0.593	0.495	0.395	0.109
"I have a lot of back pain due to two injuries. Out of everything I've taken I really like this. There is only 2 complaints: first it's expensive and second I have trouble sleeping when I take it. It wakes me up so taking it at night is not a good idea for me."	8	-0.396	0.945	0.0521	0.003
"It helped with pain, because I didn't care if I still hurt. It also interfered with my sleep if I took it before bedtime. It made my mind race and I found it hard to sleep. Weird dreams, too. During the day it is fine, but not if you need to think clearly. Take the pill and take the day off."	9	0.502	0.417	0.405	0.178
"I'd say that Ponstel was god's gift to me for treating my migraines in long term.Very good!"	10	-0.980	0.444	0.474	0.0817

IOD_mini_project3 - Sentiment Analysis

Hugging face - roBERTa (Robustly Optimized BERT Pre-training Approach):



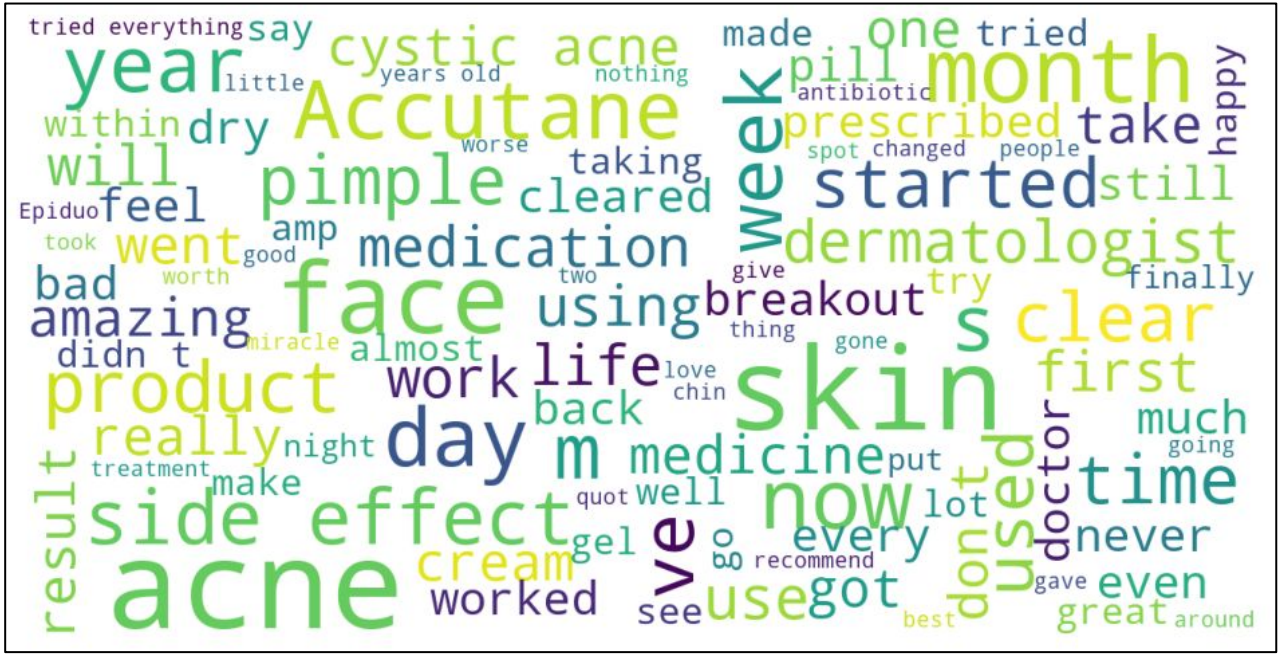




IOD_mini_project3 - Sentiment Analysis

Determine useful keywords using Word-Cloud:

- Top 25 percentile (**useful count > 20**) and **rating = 10**



IOD_mini_project3 - Sentiment Analysis

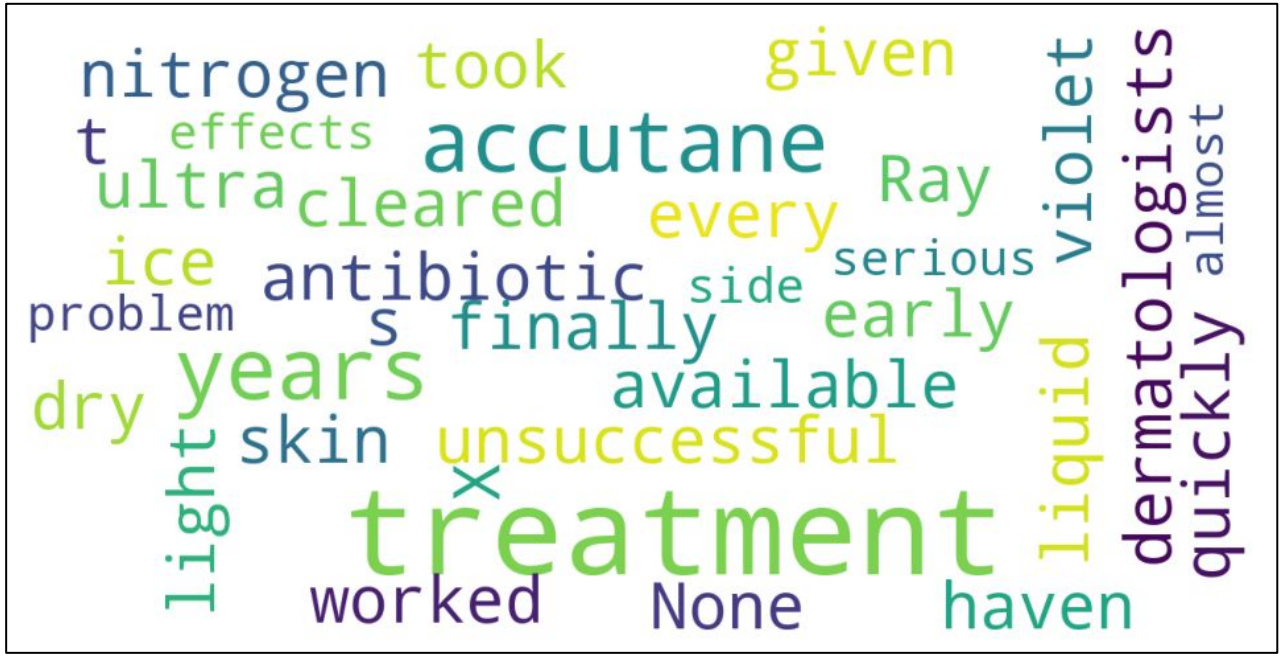
Determine useful keywords using Word-Cloud:

- Top 25 percentile (**useful count > 20**) and **rating = 1**

IOD_mini_project3 - Sentiment Analysis

Determine useful keywords using Word-Cloud:

- Top rated useful review (**useful count = 173**) and **rating = 10**



IOD_mini_project3

Findings and enhancements:

1. Classification use case:
    a. Beneficial to the medical field where predictions can be used as a second opinion to diagnosis
    b. Similar scores for different classifiers (We can look into computational resource spending as a form of optimization)
    c. Simplified problem with only 5 conditions (Use a more powerful model e.g. Neural Networks)
2. Sentiment Analysis:
    a. Determine if review is positive, neutral, or negative at a glance
        i. Validate if scoring is accurate (1-10 stars) (Useful for marketing strategy, e.g. weed out “trolling” comments)
        ii. If no scoring is available, we can determine the sentiment of the review and choose to act on it (If largely positive - reinforce marketing around it, if largely negative - improve on the drug)
    b. Analyze keywords
        i. Determine keywords that are helpful to users
    c. Use better pre-trained models for analysis (Some models are better estimators for sentiment intensity, like hugging face roBERTa)