**About this Project:**

**Given the importance of proper drugs to cure medical health our team set out to use machine**

**learning and natural language processing in an attempt to identify sentiment from the review and rating provided by the user for various medicine.**

**Team members: Nilam Roy and Indu Chandrasekharan**

**Dataset:**

**The dataset used in this project was originally published on the UCI Machine Learning repository.** "The UCI ML Drug Review dataset provides patient reviews on specific drugs along with related conditions and a 10-star patient rating system reflecting overall patient satisfaction. The data was obtained by crawling online pharmaceutical review sites. This data was published in a study on sentiment analysis of drug experience over multiple facets, ex. sentiments learned on specific aspects such as effectiveness and side effects. This dataset is collected for the period of 10 years from 2008 to 2017.

**Methodology:**

The data taken from **UCI Machine Learning repository** and stored in AWS.RDS and S3 Buckets are created to store the Big Data. The csv files are imported in Spark for cleaning purpose. After cleaning process is done the cleaned data is again loaded in RDS. Performed Vader Sentiment analysis on review. A training set of data was created separately from the test set. The training data was fitted and transformed using Python Machine Learning Pipeline that created features based on TF-IDF analysis and sentiment was vectorized using Vader Sentiment analysis.

Using the newly transformed data, a Support Vector Machine classifier model is trained. This model was used to predict whether a given submission from the test set was capturing sentiment properly or not.Accuracy was determined to be ~93%.

Tools used

Data Collection:

Python, Spark, Requests and Pandas.

Analysis:

Python, PySpark, VADER Sentiment Analysis and Pandas.

Visualization:

Flask, Wordcloud, HTML,Bootstap,CSS,Tableau

**Limitations:**

**Below are some of the challenges and limitations we came across in our project.**

* **Prediction accuracy is ~93%**
* **The model is ignorant of past health condition**
* **As review is based on drug so model is biased predicting only this condition**
* **The sentiment is categorized as only positive or negative so model can’t predict neutral comments.**

**Next Steps:**

**Below are the ideas what additional analysis can be done.**

**Predicting the best medicine based on the condition**

**Resources:**

* **Many thanks to UCI for our training data.**
* **Medium.com**