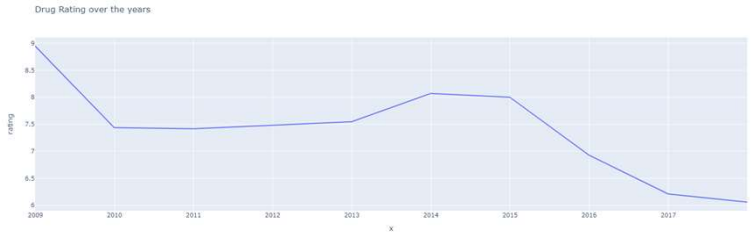
Sentiment Analysis and Condition Prediction on Drug Reviews

Team Members: Sharvil Turbadkar , Yeshwant Reddy, Akshay Bhala

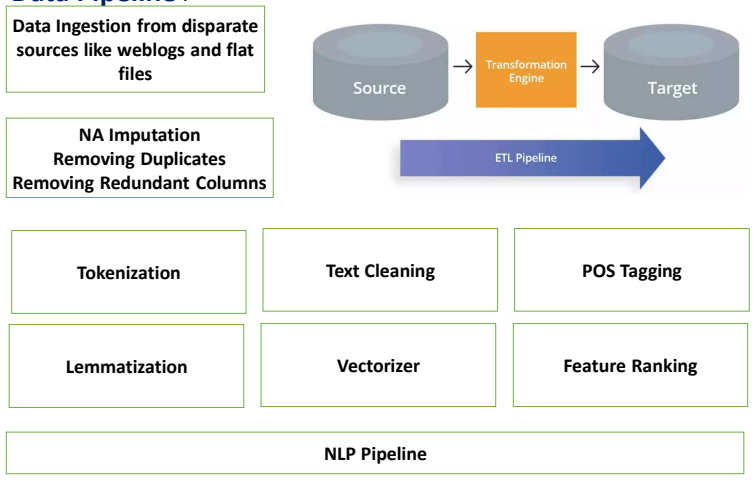
Problem : Since 2009 the reviews for Upstate Universal hospital have taken a hit. Despite strides in medical technology patient ratings for drugs are plummeting. The Pharma team is not well equipped with patient insights and to solidify customer support the upstate hospital is



Approach: My Experiences as the technical head at the Data Science Club sculpted me to be adept in Spark and Applied Natural Language Processing Techniques,.

- We aimed to research methodologies that would equip the pharma team at the upstate universal hospital with essential recommendations on how they could enhance the patient experience
- We performed sentimental analysis by crawling user data online and predicting the condition of the patient based on his reviews
- Using Spark and Natural Language Processing we were able to process huge volumes of data and formulate insights that would equip the pharma team with adept knowledge
- To predict drug effectiveness, we had to curate customer drug and review data by building ETL Pipelines would equip pharma teams with a strategy to reach out to patients

Data Pipeline :



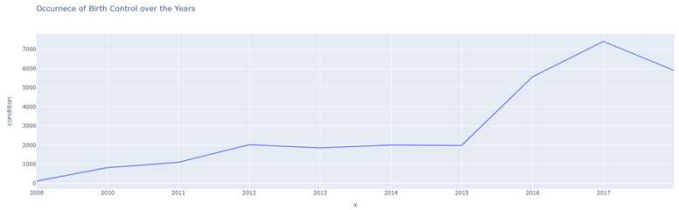
Data Description :The data has 161297 reviews. Furthermore, we had features such as drug Name , condition (name of the condition), rating (10-star patient rating), date (date of review entry), usefulCount (number of users who found the review useful). The dataset is further split into train and test sets. The dataset comprises of 885 Unique number of conditions and 3436 unique drugs. The average number of reviews per drug is 58.86.

MODEL	DESCRIPTION	EVALUATION
Naïve Bayes	Predicting patient condition	F-1 Score
Support Vector Machine	Predicting patient condition	F-1 Score
Naïve Bayes	Predicting drug review	Recall
Support Vector Machine	Predicting drug review	Recall

Result:

- Best Minimum Document Frequency of 3 gives highest f1 score
- Snowball Stemmer works best
- Bernoulli Naive Bayes gives the best results for Boolean vectorizer and ngram(1,3) after using cross-validation
- Multinomial Naive Bayes gives the best results for TF vectorizer and ngram(1,3) after using lemmatizer
- Linear SVC gives the highest f1 score of 89% after using TF vectorizer with ngram(1,3)
- Linear SVC with term frequency input and ngram range of (1,3)

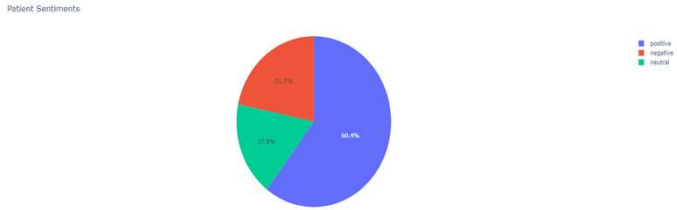
MODEL	ngrams	Precision	Recall	F1-score
Bernoulli NB	(1,3)	76.15	78.05	76.63
Multinomial NB	(1,3)	78.22	79.84	78.94
SVM	(1,3)	74.53	75.60	75.00



Insights

1)Drug ratings have been falling for the past 10 yearsThe prominent condition in 2009 is Pain while in 2017 it is Birth Control and Depression.
2) The 2008 financial recession could be one of the reasons why middle-aged adults suffer the most and need a vaccine
3)There is a steady fall in drug ratings as drugs for both birth control and depression are not alleviating the patient crisis
4)We can also conclude that abortion rates have increased dramatically in the past decade.

Patient Sentiments



Error Analysis

To analyze and assess why certain reviews were incorrectly classified we performed error analysis

Reasons why positive reviews were classified as negative

- Negative emotions associated with Side effects of the drug overpower the real meaning of the sentence

Reasons why negative reviews were classified as positive

- Underlying implicit meanings like fear which SVM was not able to classify correctly
- SVM was not able to successfully identify sarcasm within reviews