



# **Fortune Medical Associates (Predicting Drug Results)**

---

*Yamuna Umapathy*

(Data Scientist)

NLP Project

# Agenda

---

- Business Objective
- Data Understanding
- EDA (Visualization)
- Models
- Topic Modeling
- Conclusion





## **Business Objective:**

Our Stakeholder wants to predict Drug results based on patient's reviews.

- Positive Effect (Patient sent positive rating review)
- Negative (No) Effect (Patient sent negative rating review)

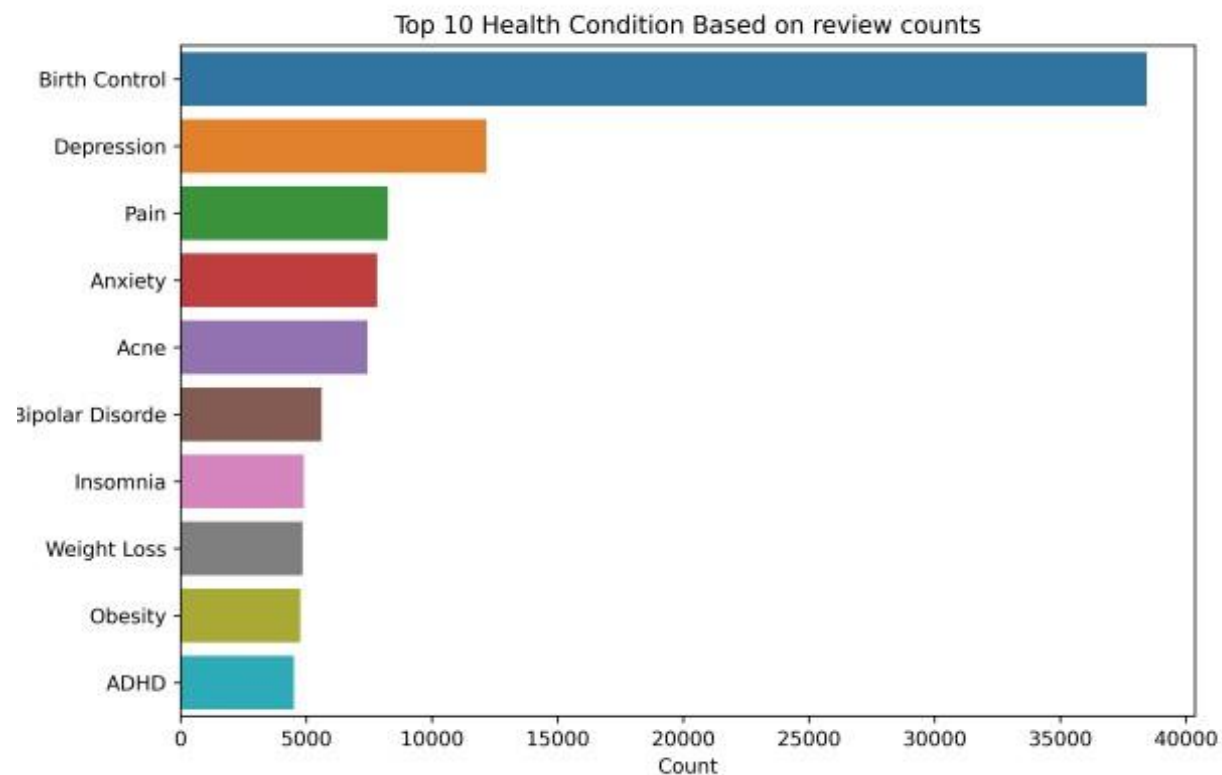
Our stakeholder's Main focus was to take care of Negative reviews, so they can follow up with patient to diagnose why Drug failed to work, which improve their Business reviews.

# Data Understanding

---

- Dataset comes from UCI.edu website  
<https://archive.ics.uci.edu/dataset/462/drug+review+dataset+drugs+com>
- Dataset has 215K rows, 6 columns
- Dropped 1194 rows for missingness, clean dataset.
- Columns: ID, drug name, condition, review, rating, date, useful count.

Based on review counts



**Target** : Column 'rating' (0 or 1)

**X** : Column 'review'(text data with 1-15 lines)

**Metrics** : Accuracy & F1 score(Balanced dataset)

Steps:

- Encoded Target as 0 for ratings (1-5) and 1 for ratings(6-10).
- Using class Text Preprocessor for cleaning X column 'review' .
  - Lowercasing, Removing stop words, punctuations, special signs.
  - Tokenization, pos tagging, lemmatize or stemming.
  - Tfidf Vectorizer or Count Vectorizer

## Data Preprocessing Target & Metrics

---



# Models (Baseline, Randomized Search CV)

## Baseline :

### Random Forest: (Test Results)

Accuracy : 65.51  
F1 Score : 65.01  
Precision : 64.08  
Recall : 65.98

## Randomized Search CV (Test Results)

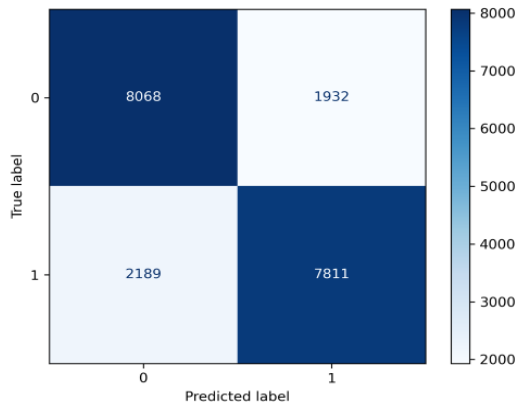
Model	Accuracy	F1 Score	Precision	Recall
Naïve Bayes	66.17	67.14	65.27	69.12
X G Boost	71.52	71.28	71.90	70.67
Random Forest	76.27	76.09	76.67	75.52

# Tensor Flow

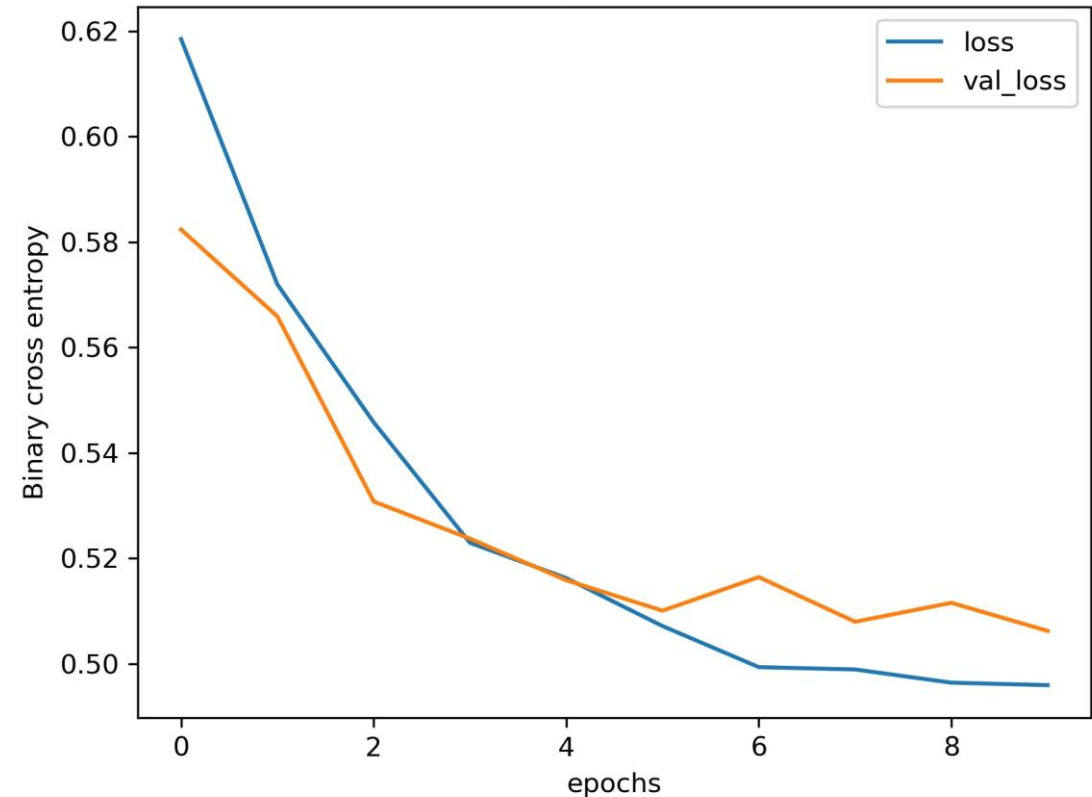
```
# defining model & hidden layers
reg = l2(3e-3)
n_features = (30141, )
tensor_med_model = Sequential()
tensor_med_model.add(Dense(32, activation='tanh', kernel_regularizer = reg, input_shape = (n_features)))
tensor_med_model.add(Dense(16, activation='tanh', kernel_regularizer = reg))
tensor_med_model.add(Dropout(0.2)) # Dropout after the first layer
tensor_med_model.add(Dense(10, activation='tanh', kernel_regularizer = reg))
tensor_med_model.add(Dropout(0.2)) # Dropout after the third layer
tensor_med_model.add(Dense(8, activation='tanh', kernel_regularizer = reg))
tensor_med_model.add(Dense(1, activation = 'sigmoid'))
```

Sequential model, Binary cross entropy,  
applying Regularization, epochs =10

Test Results:  
Accuracy: 79.39



Training loss history





# Stacking

---

- **X G Boost** – Used best model got from Randomized Search CV
- **Naïve Bayes** – Simplicity, efficient, work well with text data, cost efficiency.
- **Random Forest** – Used best model from Randomized Search CV.

**Meta Model: Logistic Regression**

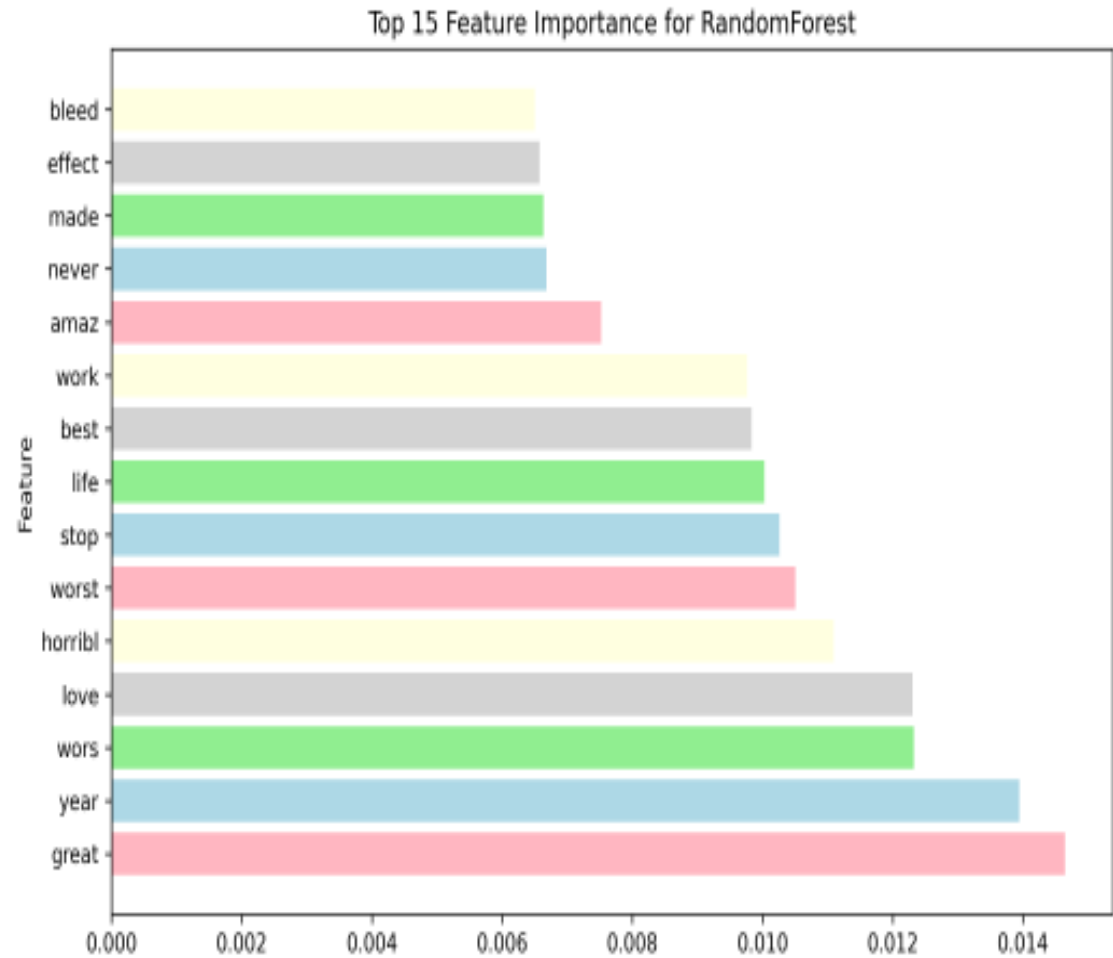
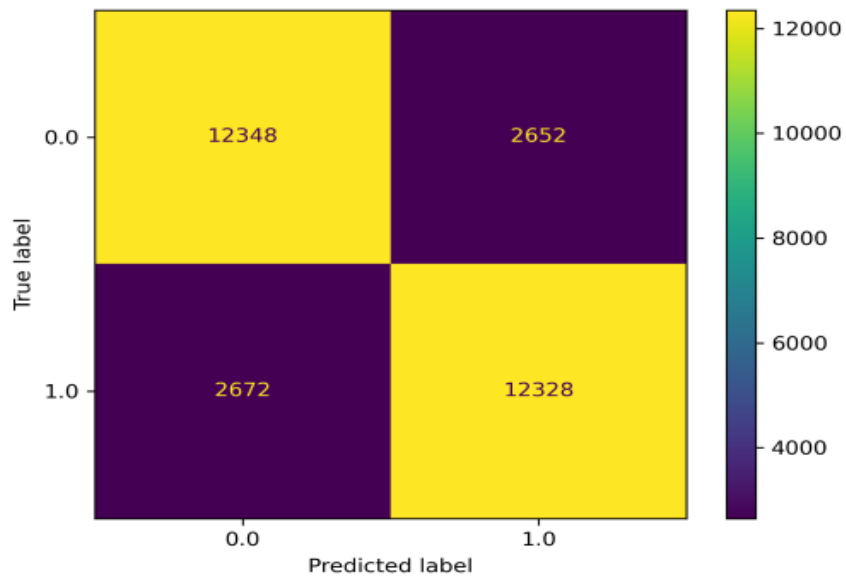
## Stacking (Best Model) Test Results:

Accuracy	82.25
F1 Score	82.24
Precision	82.18
Recall	82.29

# Stacking Visualization:

False Positive : Patient has No effect or had side effects after medication was taken.

False Negative: Patient condition improved after medication was taken.



Metrics used Accuracy and F1 score, Dataset is Balanced data.



Model	Accuracy	F1 Score
Baseline	65.51	65.01
Naïve Bayes	66.17	67.14
X G Boost	71.52	71.28
Random Forest	76.27	76.09
Tensor Flow	79.33	
Stacking	82.25	82.24

# Model Results

# Topic Modeling NMF

Topic modeling is a statistical modeling technique used to identify latent topics or themes within a collection of documents.

Using Non-Negative Matrix factorization model, and choosing n-components as 3, derived words associated with three topics.

$X \sim WH$

W – Importance of each token in fitted topics.

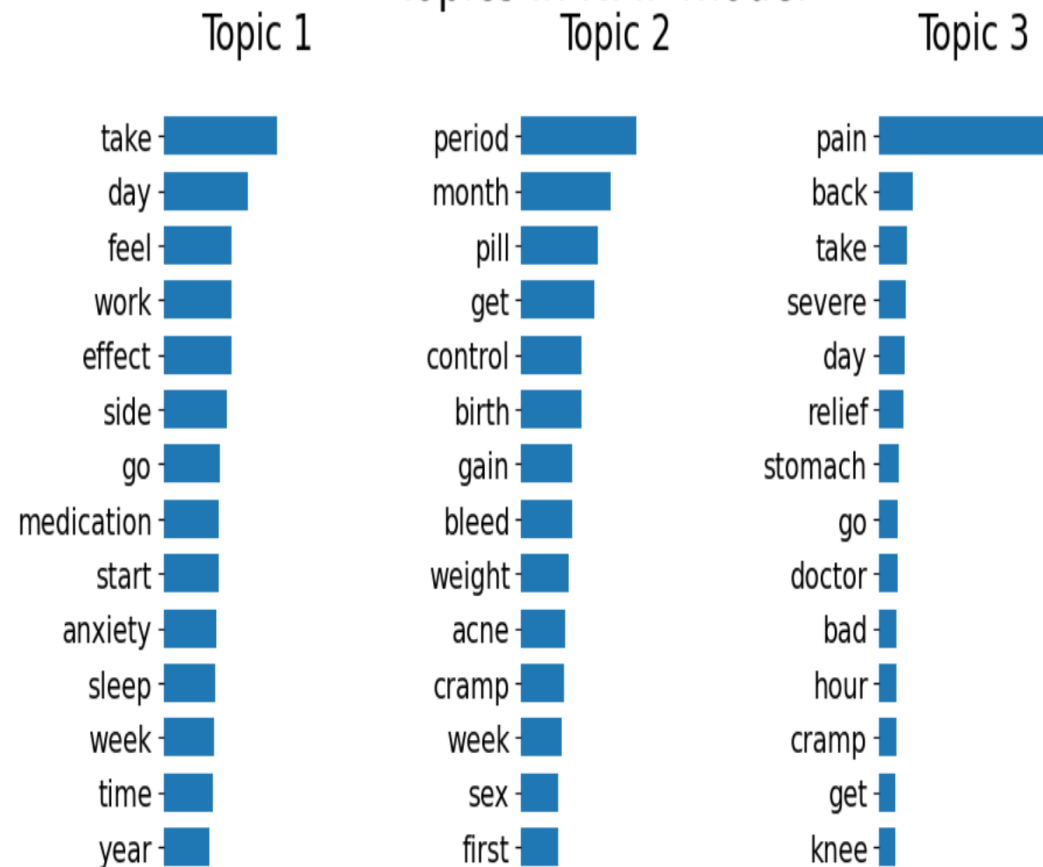
H – Weight of the fitted topic in each doc.

Topic 1 Label: Anxiety/Depression

Topic 2 Label: Birth Control/Infertility

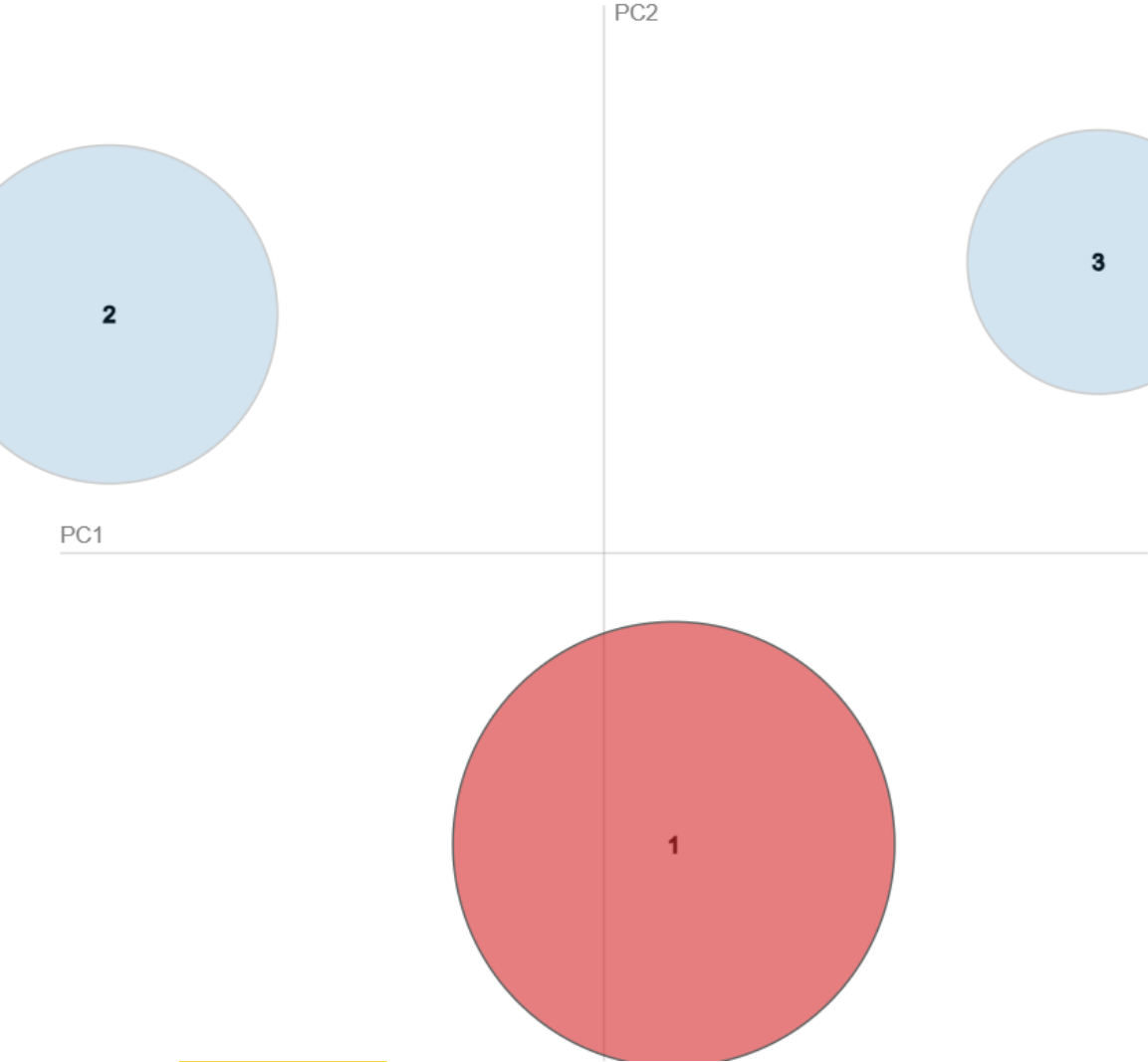
Topic 3 Label: Pain

## Topics in NMF model

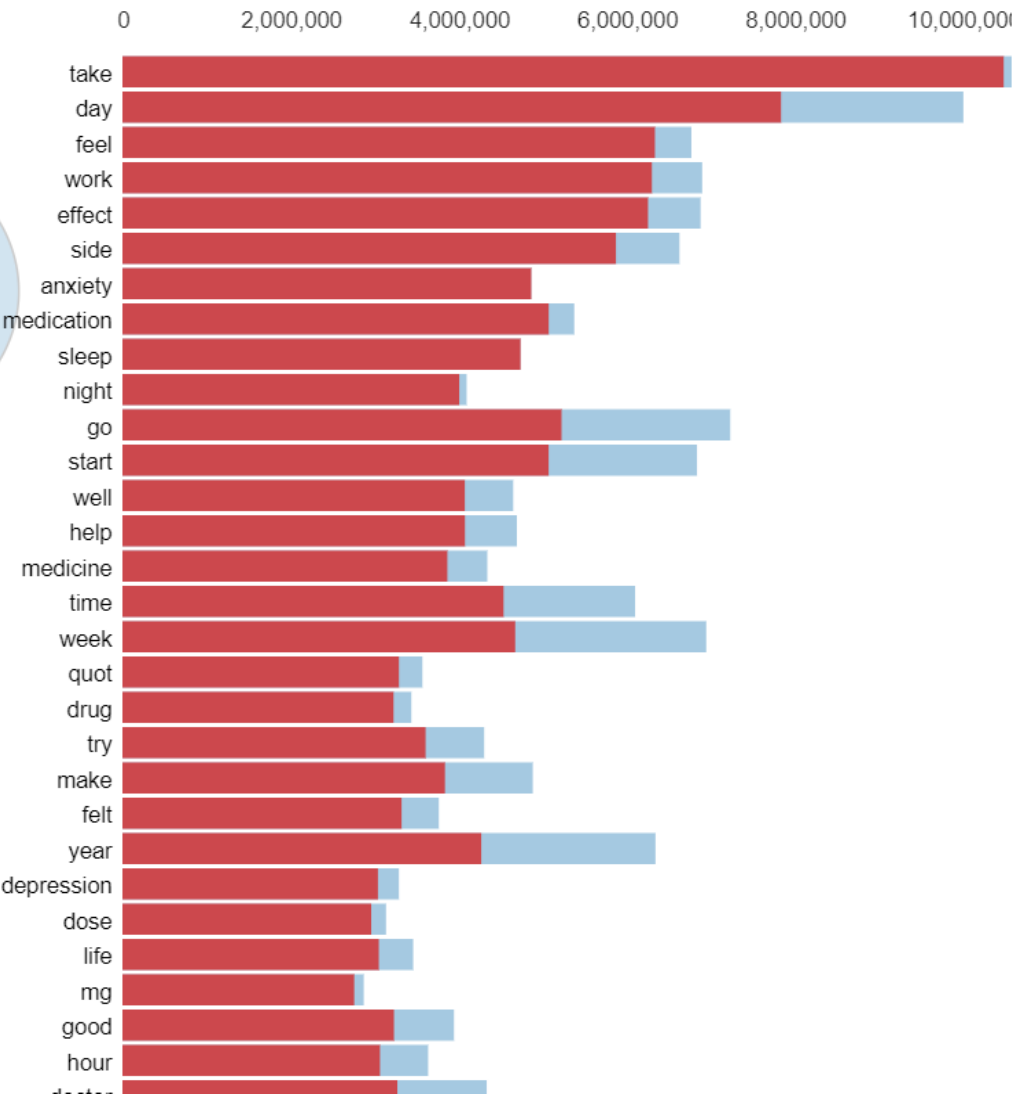


# NMF Visualization

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 1 (51.8% of to



# Conclusion:

---

1. Best Model Stacking with best accuracy 82.25 to solve our Stakeholder's problem.
2. Topic Modeling helps to visualize health conditions as labels assuming as Birth Control, Anxiety & Pain.
3. Concentrate on True Negative & False Positive results, these patients need immediate follow up to diagnose why Drug failed to work.

## Next Steps:

- For Phase 2, use Word2Vec algorithm, widely used word representation technique.
- Implement TextBlob or PySpellCheck library to correct spelling mistakes for medical terms before running next model.

# Thank you & Questions?

---

Yamuna Umapathy (NJ)

Email : [u.yamuna@gmail.com](mailto:u.yamuna@gmail.com)

Linked in: <https://www.linkedin.com/in/yamuna-Umapathy/>

