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.

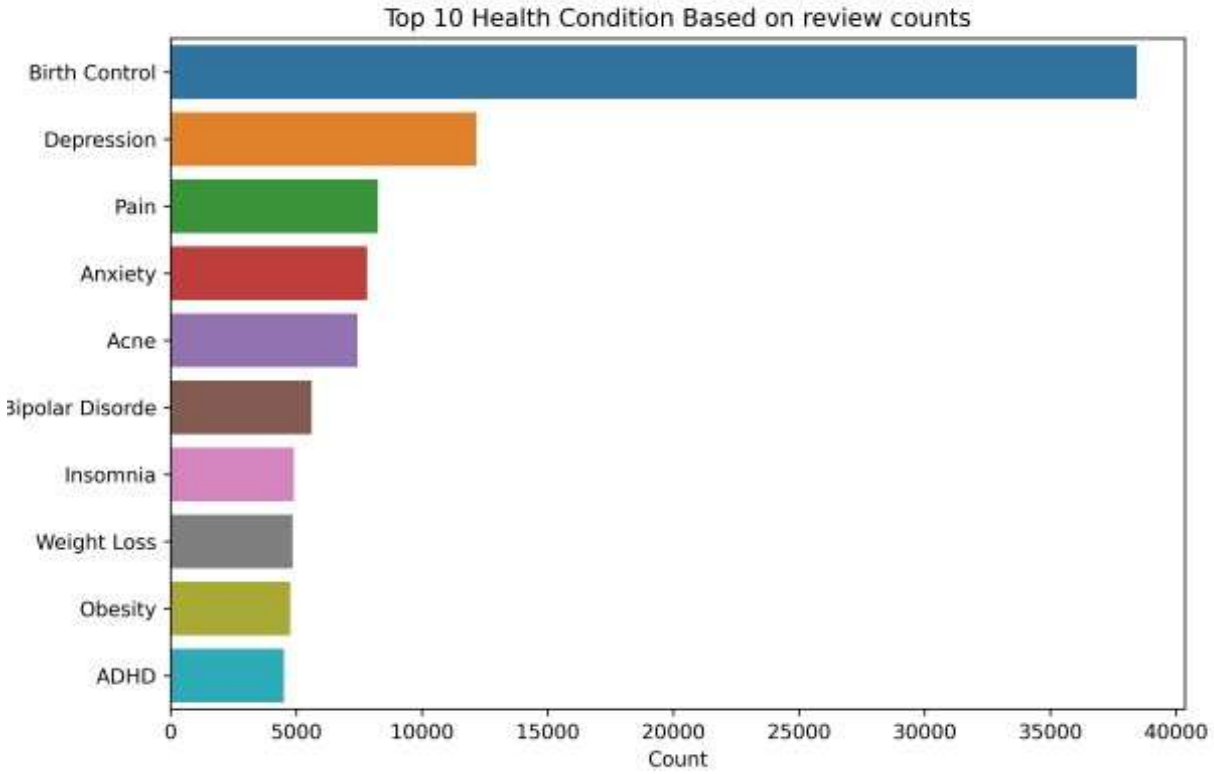
Data Understanding

- 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.
- Choose Balanced Sample size of 100K rows from dataset, encoded target 0 & 1.

EDA


WordCloud from Reviews with Positive Rating
nightmare
week
work
littleurine
month
pregnancy
role
duration
condemn
fruit
whatif
comforter
antidepressant
breakup
funeral
key
mark
give
wd
objection
hematoma
shift
poison
praise
migrain
outlet
aggression
vertabra
waffle
lance
dwindle
protien
condom
cling
lighten
abortive
lymph
prep
content
spark
relaxant
recurrence
salofalk


Top 10 Health Condition Based on review counts
Birth Control
Depression
Pain
Anxiety
Acne
Bipolar Disorde
Insomnia
Weight Loss
Obesity
ADHD
0
5000
10000
15000
20000
25000
30000
35000
40000
Count

Data Preprocessing Target & Metrics

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

Models (Baseline, Randomized Search CV)

Baseline :

Random Forest:
(Test Results)

Accuracy : 65.51

F1 Score : 65.01

Precision : 64.08

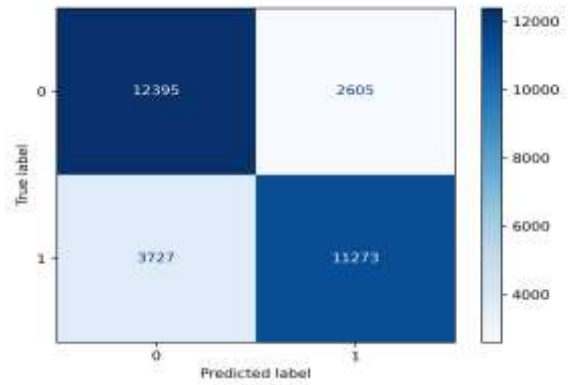
Recall : 65.98

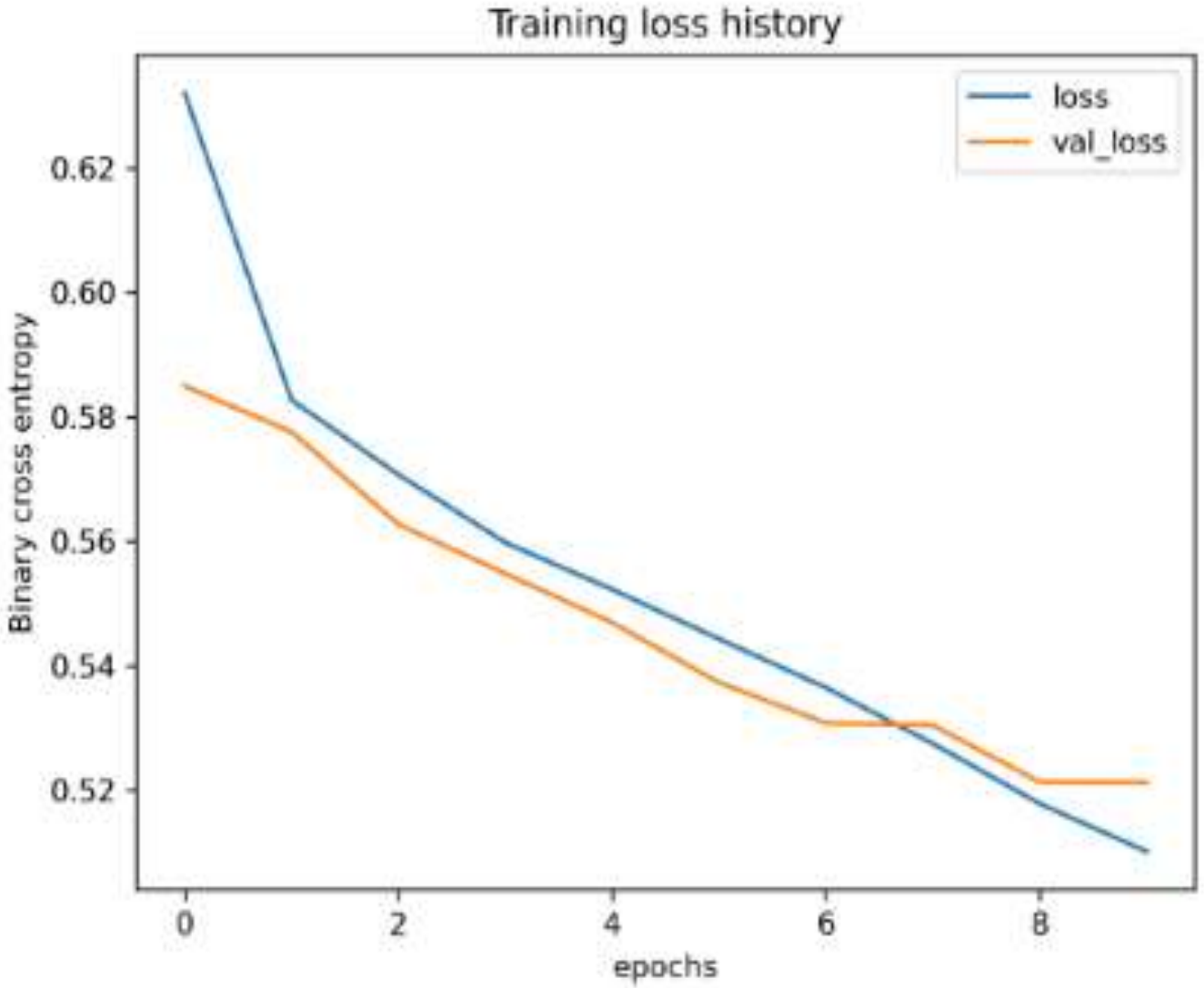
Randomized Search CV (with tuned hyperparameters):

	Random Forest (Test Results)	X G Boost (Test Results)
Accuracy :	76.27	70.83
F1 Score:	76.09	70.55
Precision:	76.67	71.23
Recall :	75.52	69.89

Tensor Flow

Sequential model, Binary cross entropy, applying Regularization, epochs =10
Test Results:
Accuracy: 78.89





Stacking

- **Linear SVC** – Efficient with Large datasets, handles high dimensional sparse data.
- **Naïve Bayes** – Simplicity, efficient, work well with text data, cost efficiency.
- **Random Forest** – Used best model predicted by Randomized Search CV.

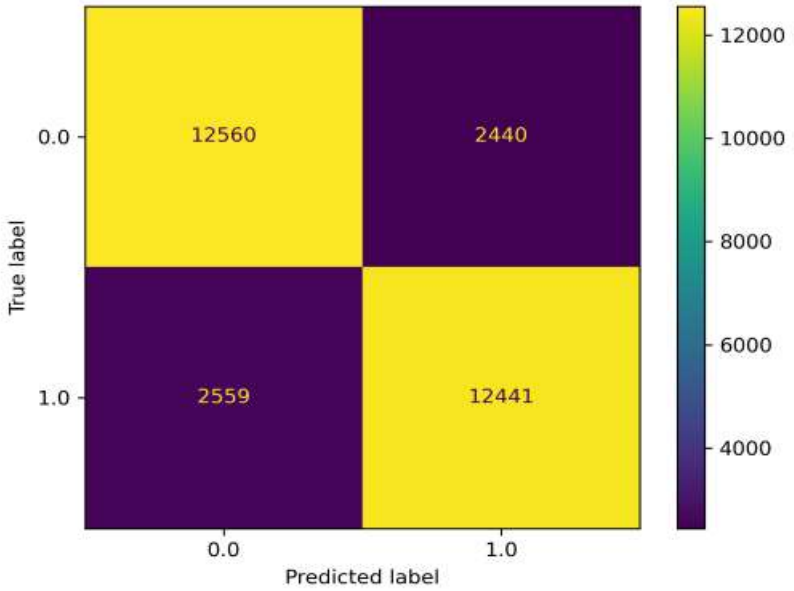
Meta Model: Logistic Regression

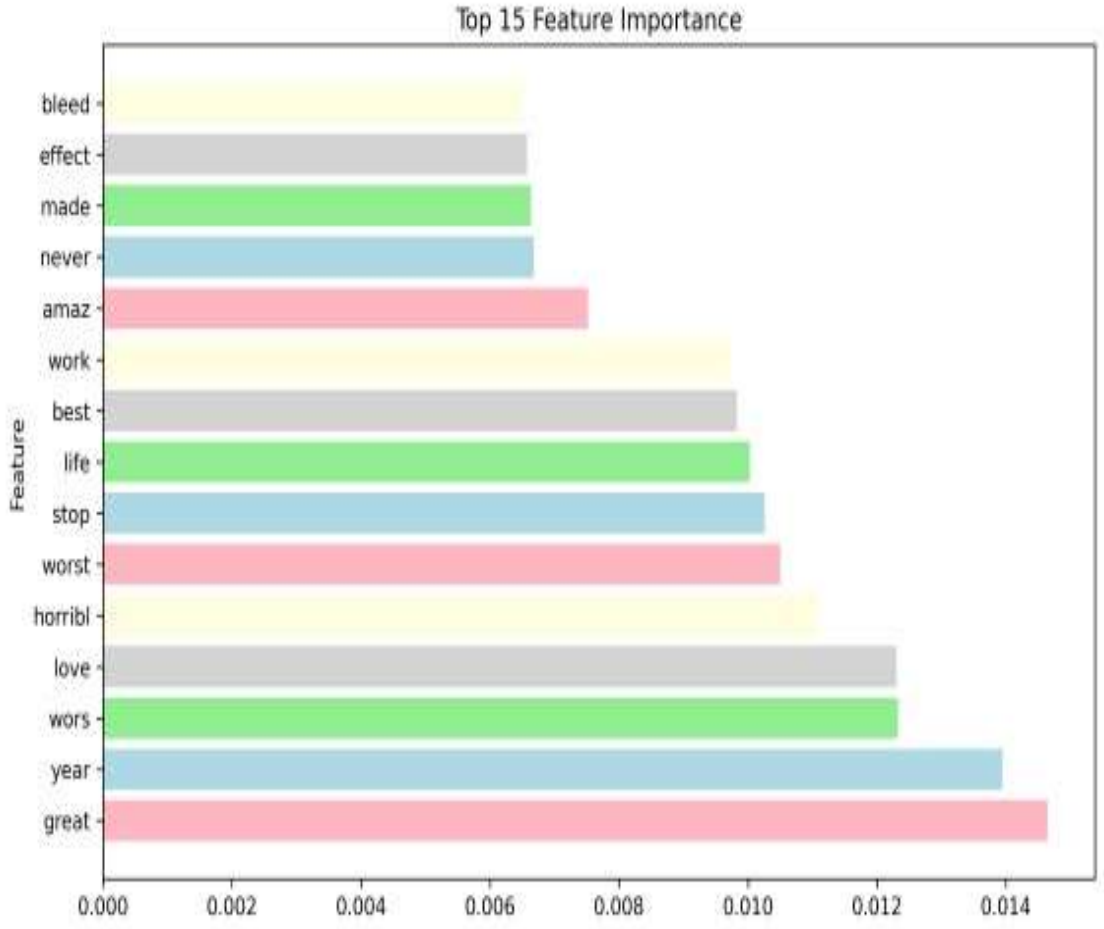
Stacking (Best Model) Test Results:

Accuracy	83.33
F1 Score	83.27
Precision	82.94
Recall	83.60

Stacking Visualization:

True Positive : Positive Effect after medication.
True Negative : Negative Effect after medication.
False Positive : Negative Effect
False Negative: Positive Effect





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

Model	Hyperparameter	Accuracy	F1 Score
Baseline Random Forest		65.5	65.0
Random Forest	Randomized Search CV	76.2	76
X G Boost	Randomized Search CV	71.6	15
Tensor flow		78.89	
Stacking(Linear SVC, Naïve Bayes, Logistic Reg)		83.36	83.27

Model Results

Topic Modeling NMF

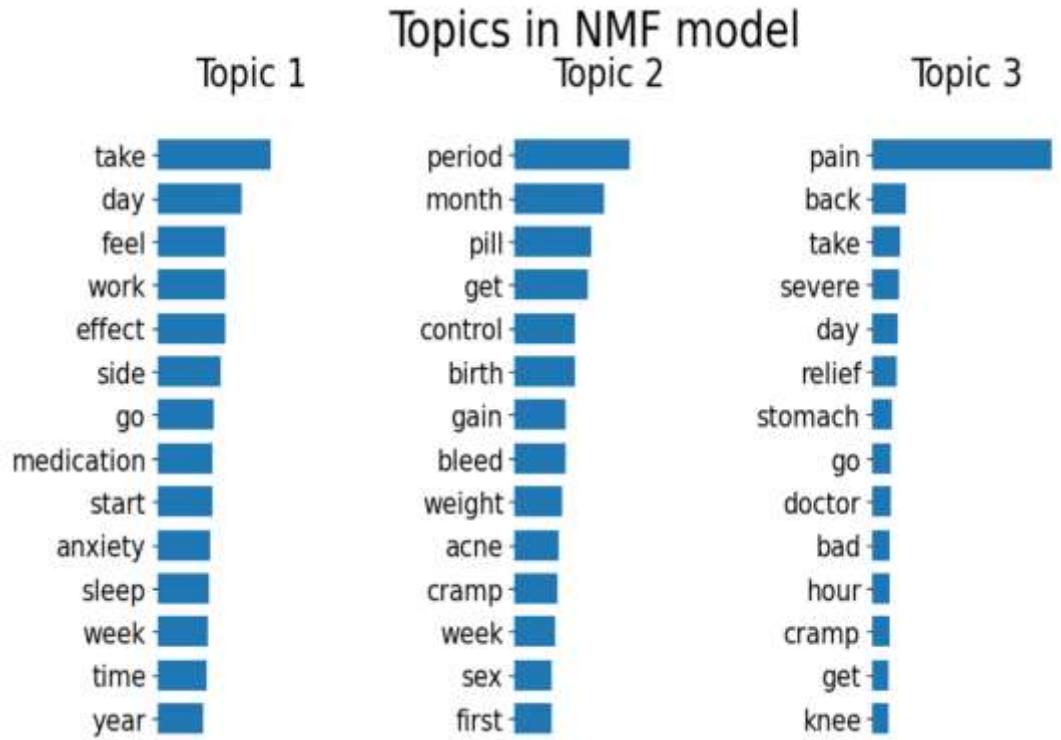
Topic modeling is a statistical modeling technique used to identify latent topics or themes within a collection of documents.

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

X ~ WH

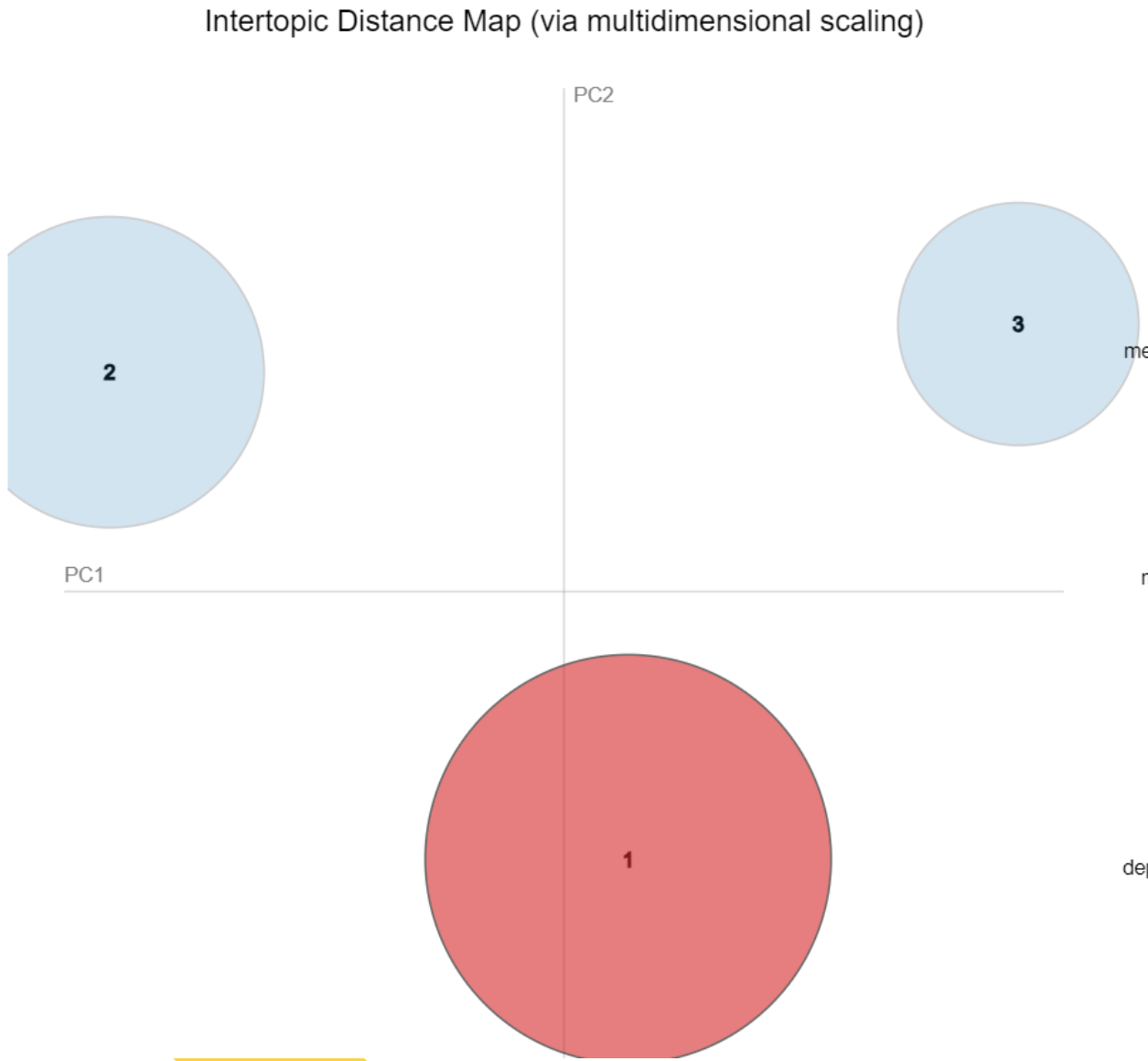
W – Importance of each token in fitted topics.

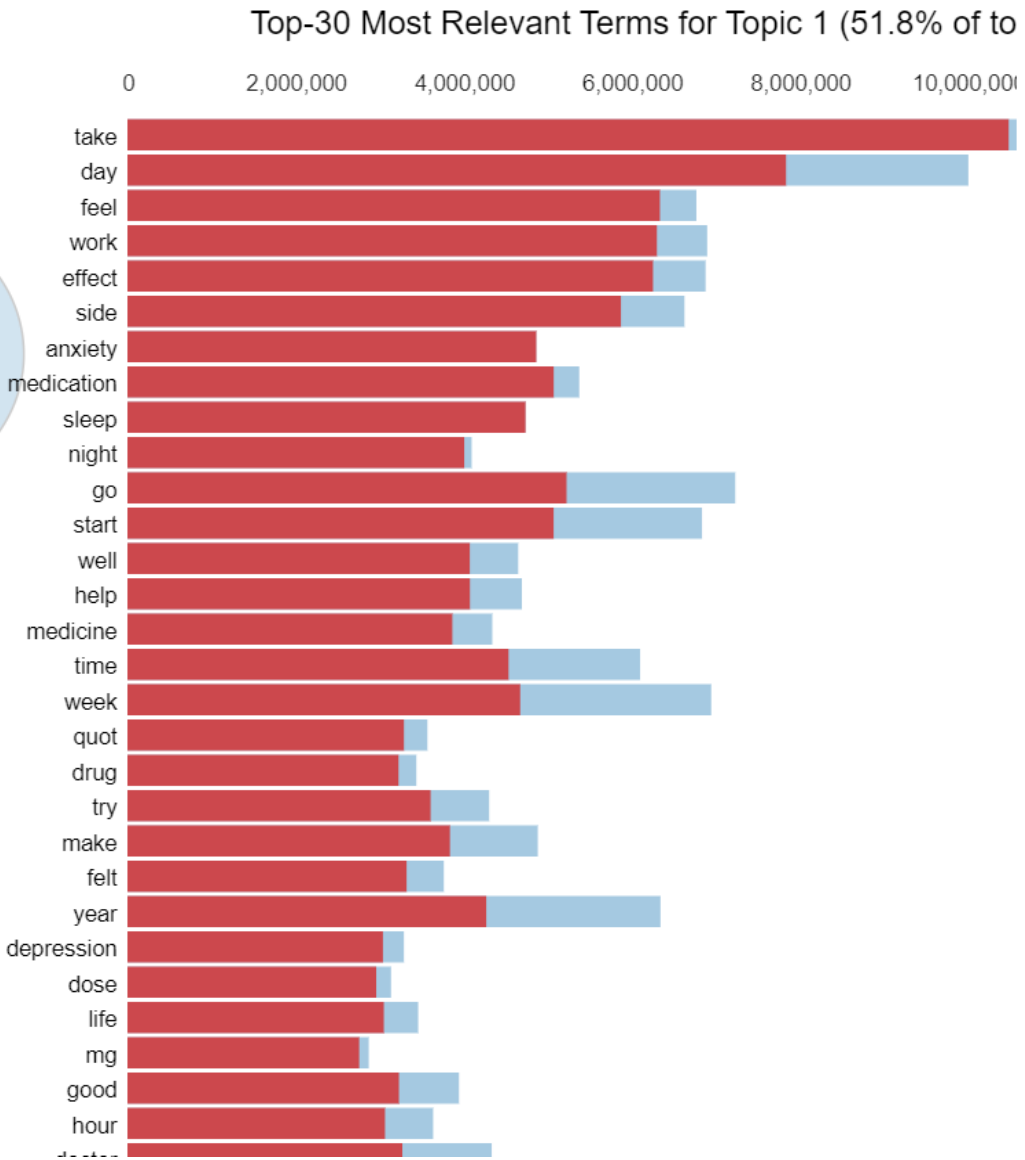
H – Weight of the fitted topic in each doc.



Html link: C:/Users/uyamu/Documents/FortuneMedical/nmf_topics.html#topic=2&lambda=1&term=

NMF Visualization





Conclusion:

1. Best Model Stacking with better accuracy 83.3 to solve our Stakeholder's problem.
2. Topic Modeling helps to visualize health conditions assuming as Birth Control, Anxiety & Pain.
3. Concentrate on True Negative & False Positive results, these patients needs immediate follow up to diagnose why Drug failed to work.

Next Steps:

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

Thank you & Questions?

Yamuna Umapathy (NJ)

Email : u.yamuna@gmail.com

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