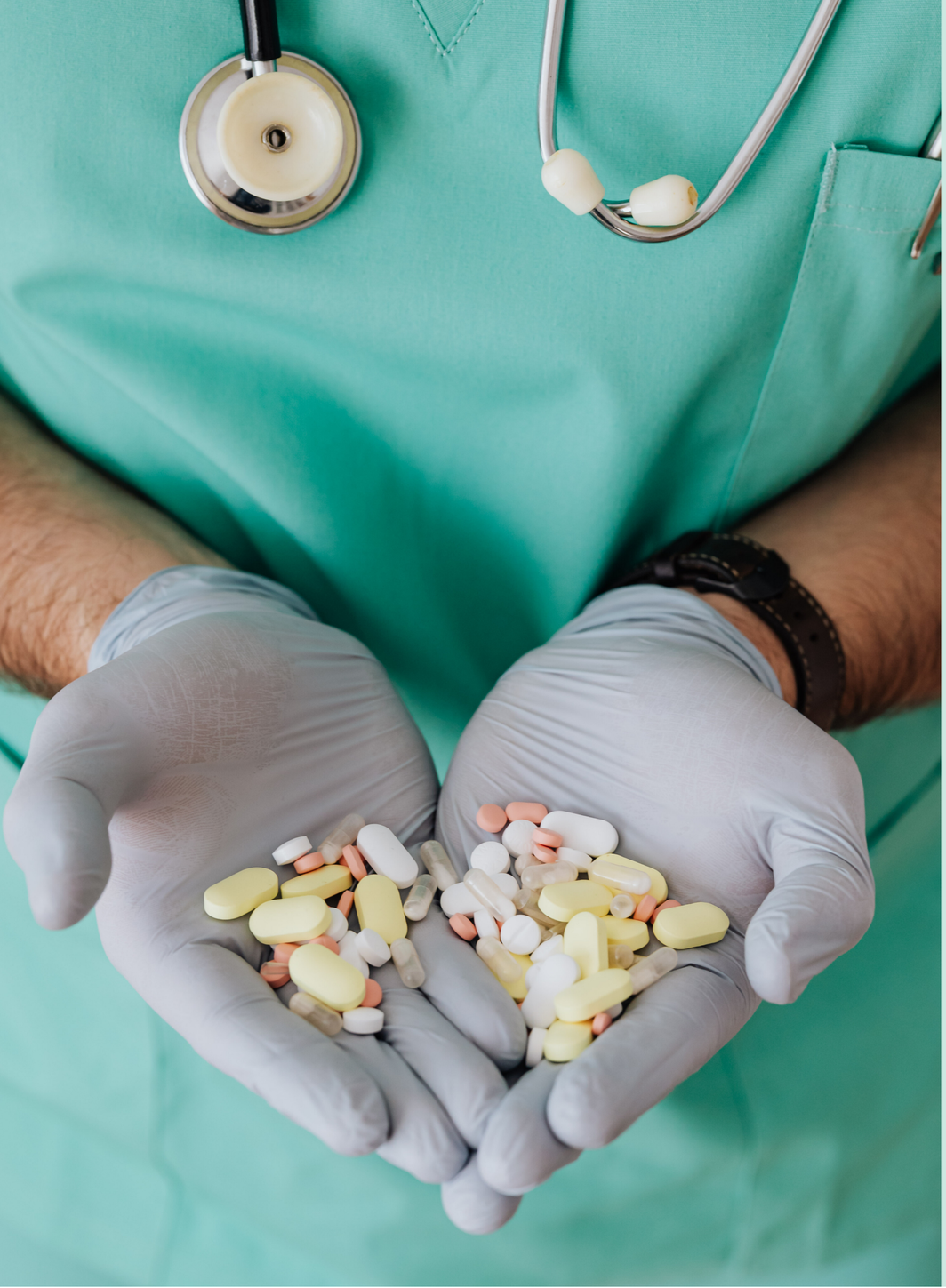


Sentiment Analysis (and beyond)

UCI Drug Dataset

Odelia Ahdout
Ironhack DA, 2022-I

Presentation Outline

Healthcare meets IT

- **This dataset: healthcare/IT Interface**
- Goal: track the safety and effectiveness of drugs over time.
- Clinical trials are limited in number of participants and timespan (e.g. Covid vaccination..).
- Through user reviews a big amount of information about the over time.

Features in the dataset

DRUG
(CATEGORICAL)

CONDITION
(CATEGORICAL)

REVIEW
(TEXT)

PREDICTOR

USEFUL_COUNTS
(NUMERICAL)

DID NOT USE
(ASK ME ABOUT IT)

RATING [1-10]
(NUMERICAL)

TARGET VARIABLE

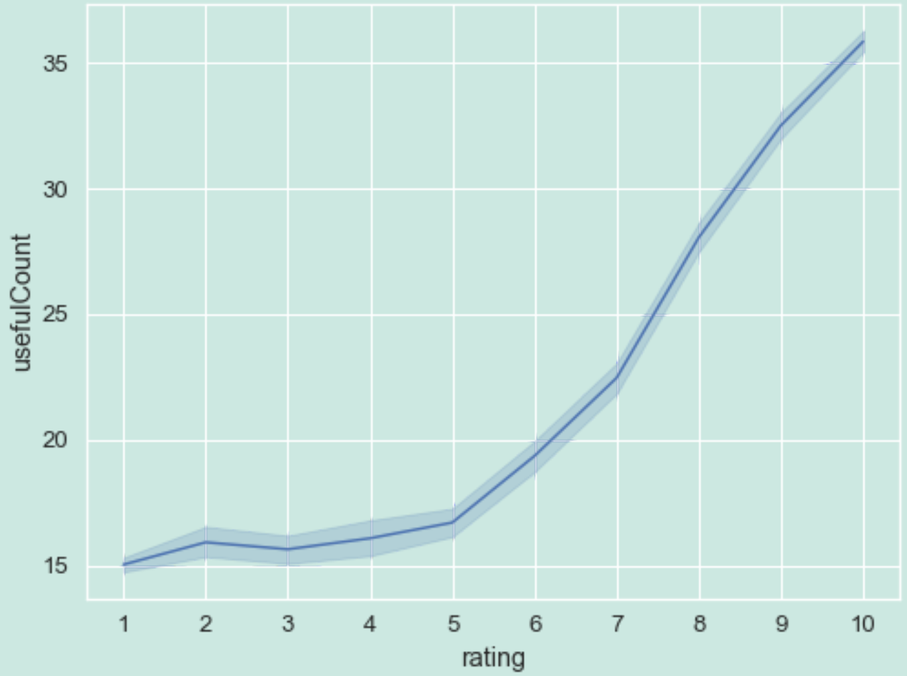
CONVERTED TO
BINARY:
1-6 > "0"
7-10 > "1"

Duplication alert!

- Scraping error in data collection caused the dataset to have a large amount of duplicates (not addressed in the research paper!).
- Duplicates = same text input repeated twice => biased model.
- After removal, dataset shrank almost by half.

Usefulness votes

Average of UsefulCount per Rating



Comments associated with positive rating tend to receive more 'useful comment' votes.

However, data is imbalanced towards positive reviews (see next slide):

- Removing less useful ones would decrease negative reviews
- i.e. **increase imbalance**.

=> No removing of records

Drug types and medical conditions in the dataset

3,199
UNIQUE DRUGS

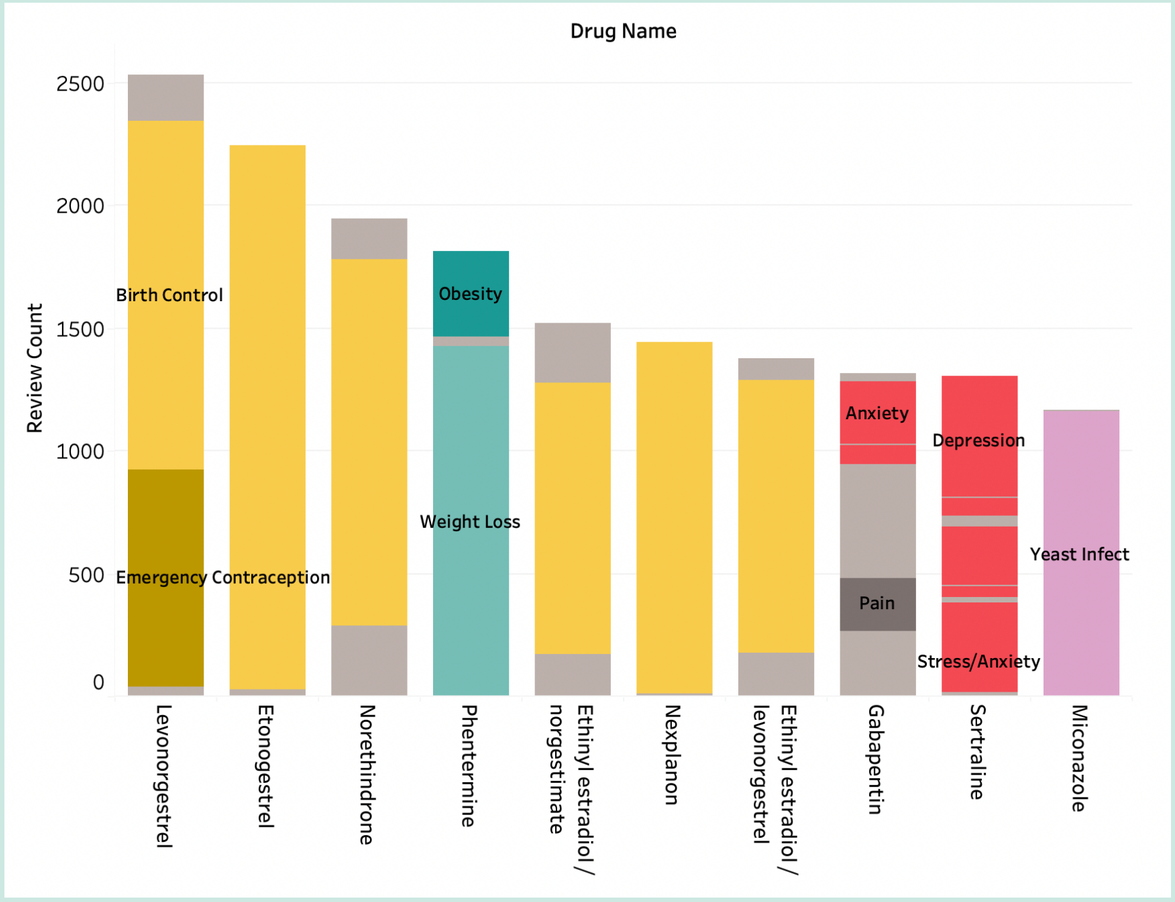
837
UNIQUE CONDITIONS

130,285
UNIQUE REVIEWS

* COLLECTED OVER A 9 YEAR TIME SPAN (2008-2017)

Drug types and medical conditions in the dataset

Top 10 reviewed drugs, with most common target Condition

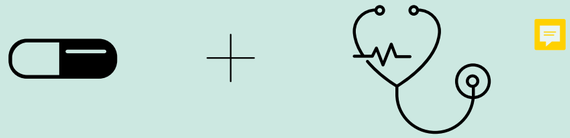


Most common conditions targeted by top 10 drugs in the dataset:

- Birth control and termination
- Mood disorders
- Weight loss

Drug recommender

The dataset allows users to find the best and worst drugs for the condition they are suffering from
press below for link:



Main predictive feature: patient reviews

POSITIVE REVIEW

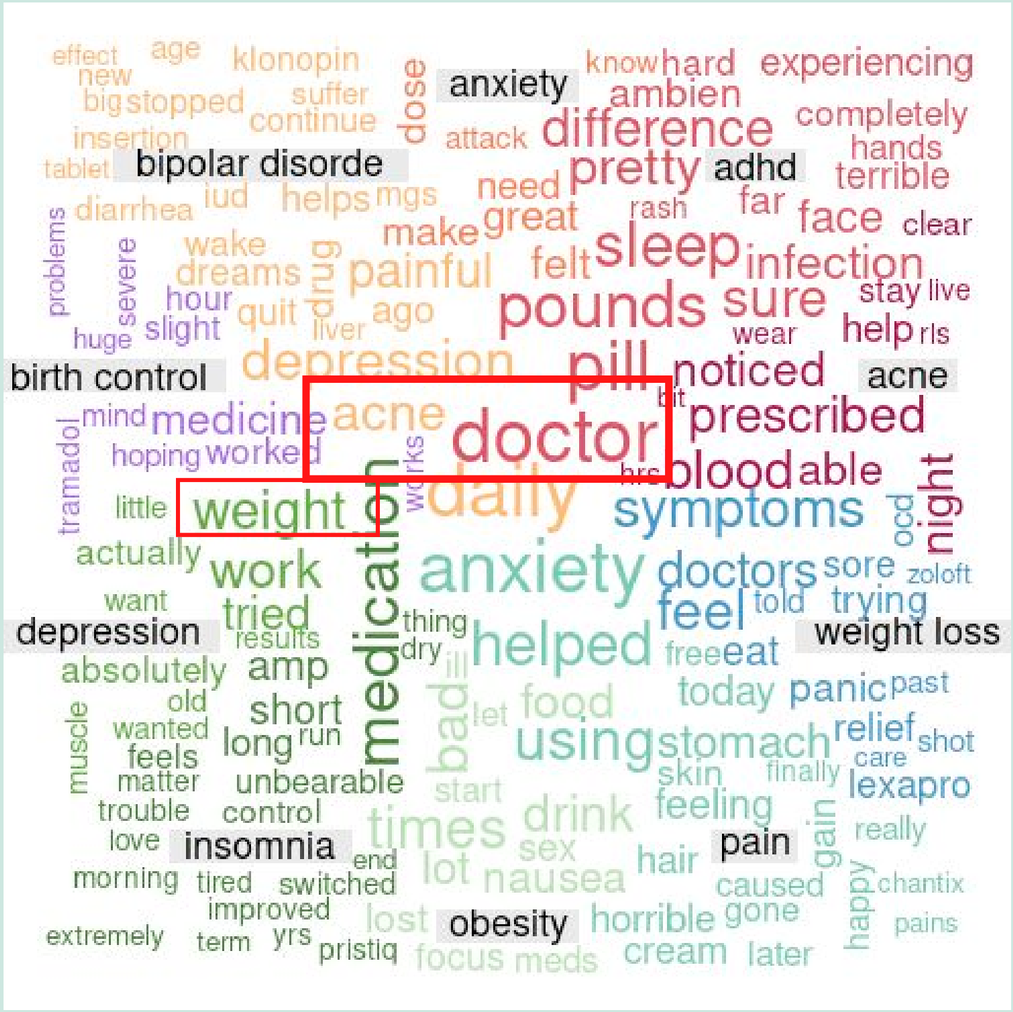
"Absolutely saved my life. I've tried numerous drugs over 15 years and this is a miracle. No side effects whatsoever. I feel better than I ever have in my life".

NEGATIVE REVIEW

"Absolutely terrible experience on Tirosint - landed in ER with chest pains, dizziness, extremely disoriented."

Word Clouds

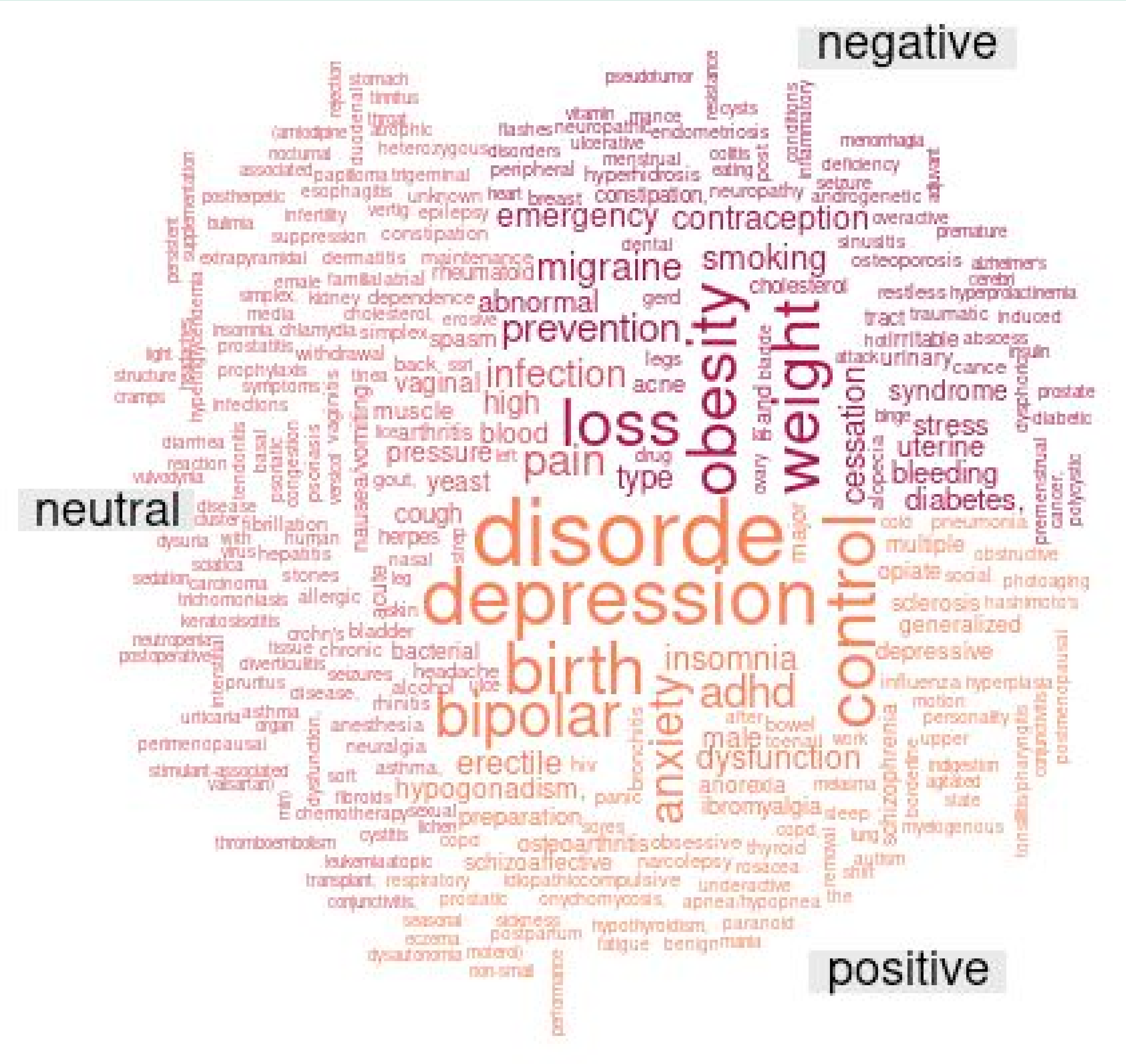
Most frequent words associated with **top 10 Conditions**



- Credit goes to Giovanni Marelli for assisting with Word Clouds.

Word Clouds

Reviews compared against word lists associated with positive / neutral / negative items

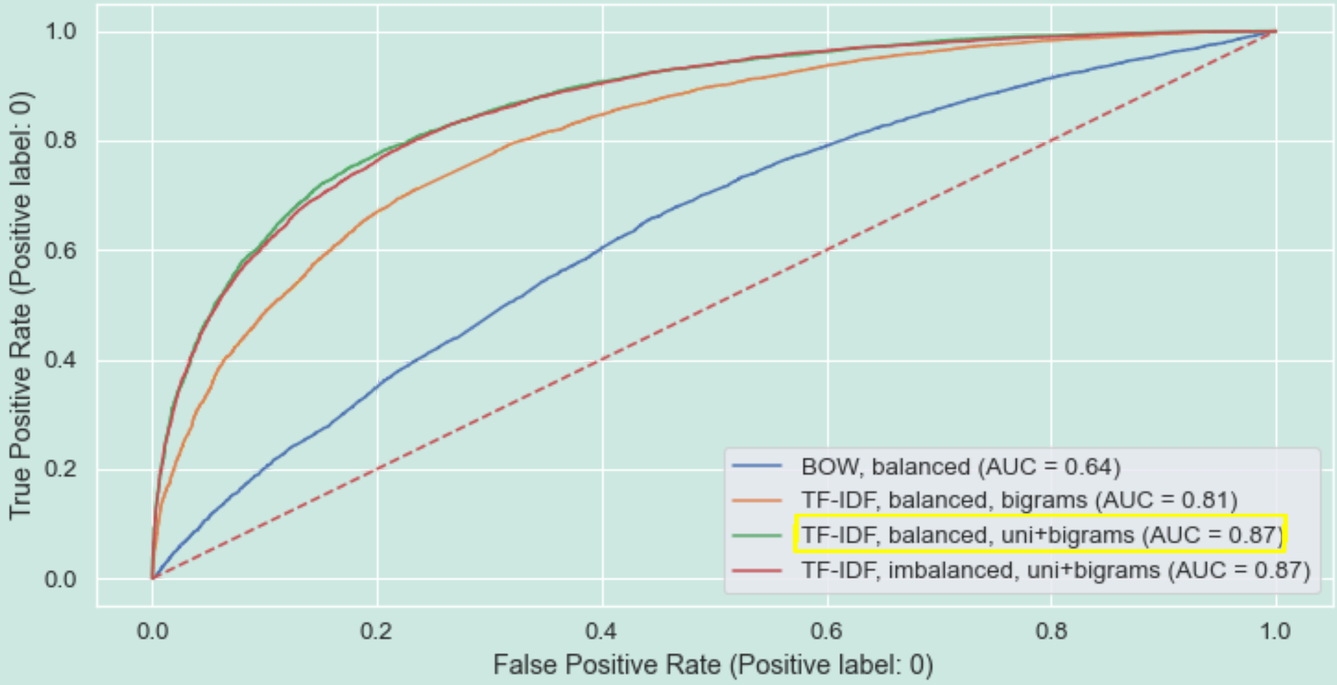


drug dataset

Model Comparison

- Using Random Forests

1. TF-IDF > BOW
2. Downsample > full sample
    - (Better recall for both categories)
3. Uni + bigrams > bigrams only



Overall Best model:

TF-IDF + uni & bigrams

[balanced sample]

Conclusions and Future Ideas

- **General**:
  - Most common conditions people review on are **birth control**, **mood disorders**, **pain**, **obesity,** and **skin conditions**.
  - The current dataset allows us to identify medical conditions which are relatively well treated (e.g. depression), and target ones which are not (e.g. weight loss, female hair loss), and require more research.

- **Sentiment Analysis**:
  - Advantage: allows a much more fine grained assessment of drugs and medical conditions, discovering new associations previously unexpected.
  - Would improve with defining a 'neutral' category (did not make it to incorporate in actual model).
  - Use models which allow more semantic mapping of words than BOW/TF-IDF allow (Word2Vec).

THANKS!

Appendix

Balanced

Results obtained for the TRAIN SET
==========
The Cohen's Kappa is: 0.96

	precision	recall	f1-score	support
0	0.98	0.98	0.98	34967
1	0.98	0.98	0.98	35194
accuracy			0.98	70161
macro avg	0.98	0.98	0.98	70161
weighted avg	0.98	0.98	0.98	70161

==========
Results obtained for the TEST SET
The Cohen's Kappa is: 0.57

	precision	recall	f1-score	support
0	0.79	0.79	0.79	8884
1	0.78	0.78	0.78	8657
accuracy			0.79	17541
macro avg	0.79	0.79	0.79	17541
weighted avg	0.79	0.79	0.79	17541

Imbalanced, more 1s!

Results obtained for the TRAIN SET
==========
The Cohen's Kappa is: 0.99

	precision	recall	f1-score	support
0	1.00	0.99	0.99	35026
1	0.99	1.00	1.00	69202
accuracy			1.00	104228
macro avg	1.00	0.99	0.99	104228
weighted avg	1.00	1.00	1.00	104228

==========
Results obtained for the TEST SET
The Cohen's Kappa is: 0.46

	precision	recall	f1-score	support
0	0.84	0.45	0.58	8825
1	0.77	0.96	0.85	17232
accuracy			0.78	26057
macro avg	0.81	0.70	0.72	26057
weighted avg	0.80	0.78	0.76	26057