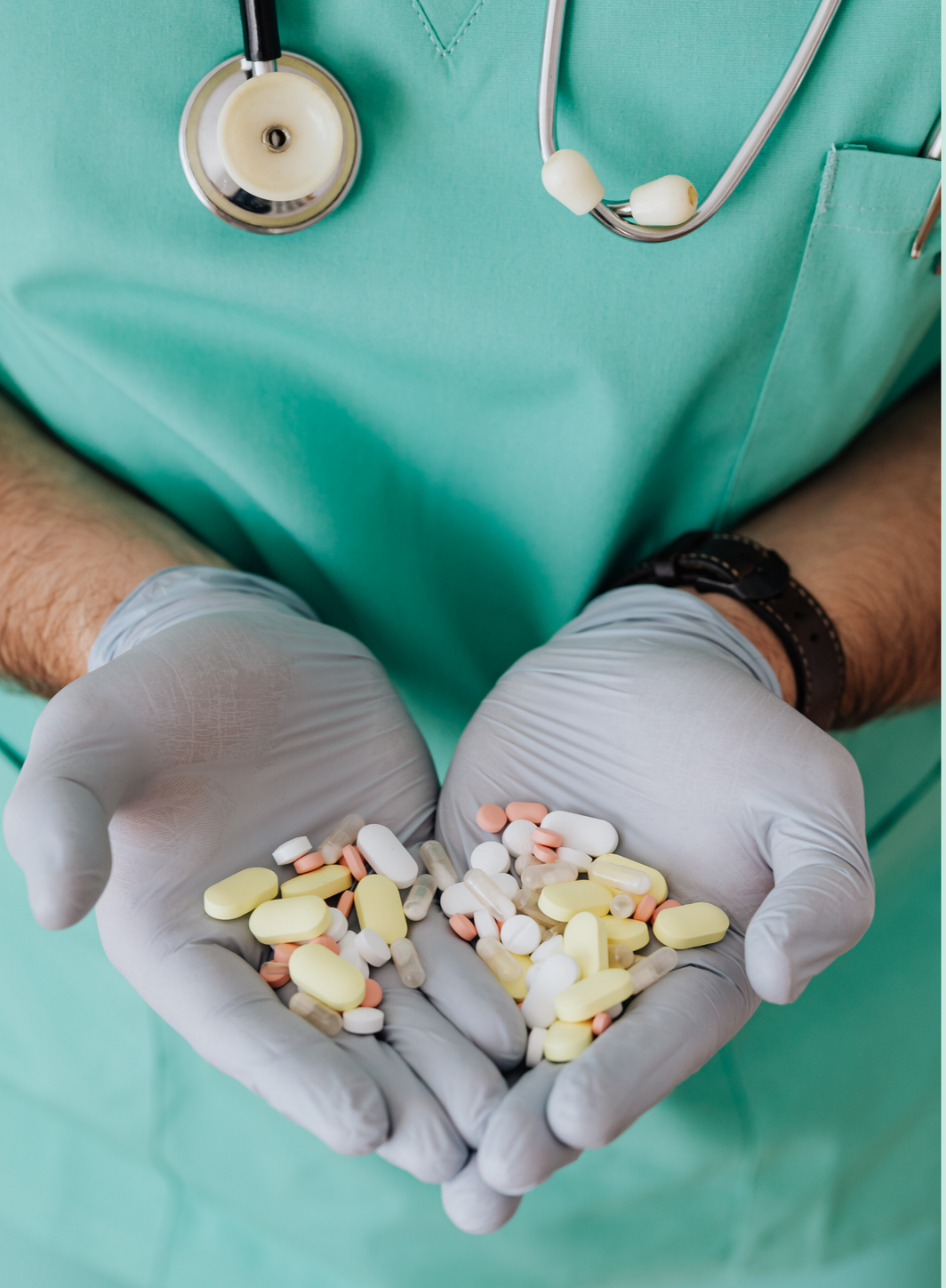


SENTIMENT ANALYSIS OF DRUG REVIEWS

UCI ML Drug Review dataset

Odelia Ahdout
Ironhack DA, 2022-I

Presentation Outline

Healthcare meets IT

- Major goals in the health and pharmaceutical industry is to ensure the effectivity and safety of drugs.
- Pharmaceutical product safety depends on clinical trials and specific test protocols.
- Such studies are typically done under strict conditions
  - Limited number of test subjects
  - Limited time span.
- User reviews as a resource offer great potential in obtaining such data.

Features in the dataset

DRUG
(CATEGORICAL)

CONDITION
(CATEGORICAL)

REVIEW
(TEXT)

PREDICTOR

USEFUL_COUNTS
(NUMERICAL)

RATING [1-10]
(NUMERICAL)

TARGET VARIABLE

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.

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."

Drug types and medical conditions in the dataset

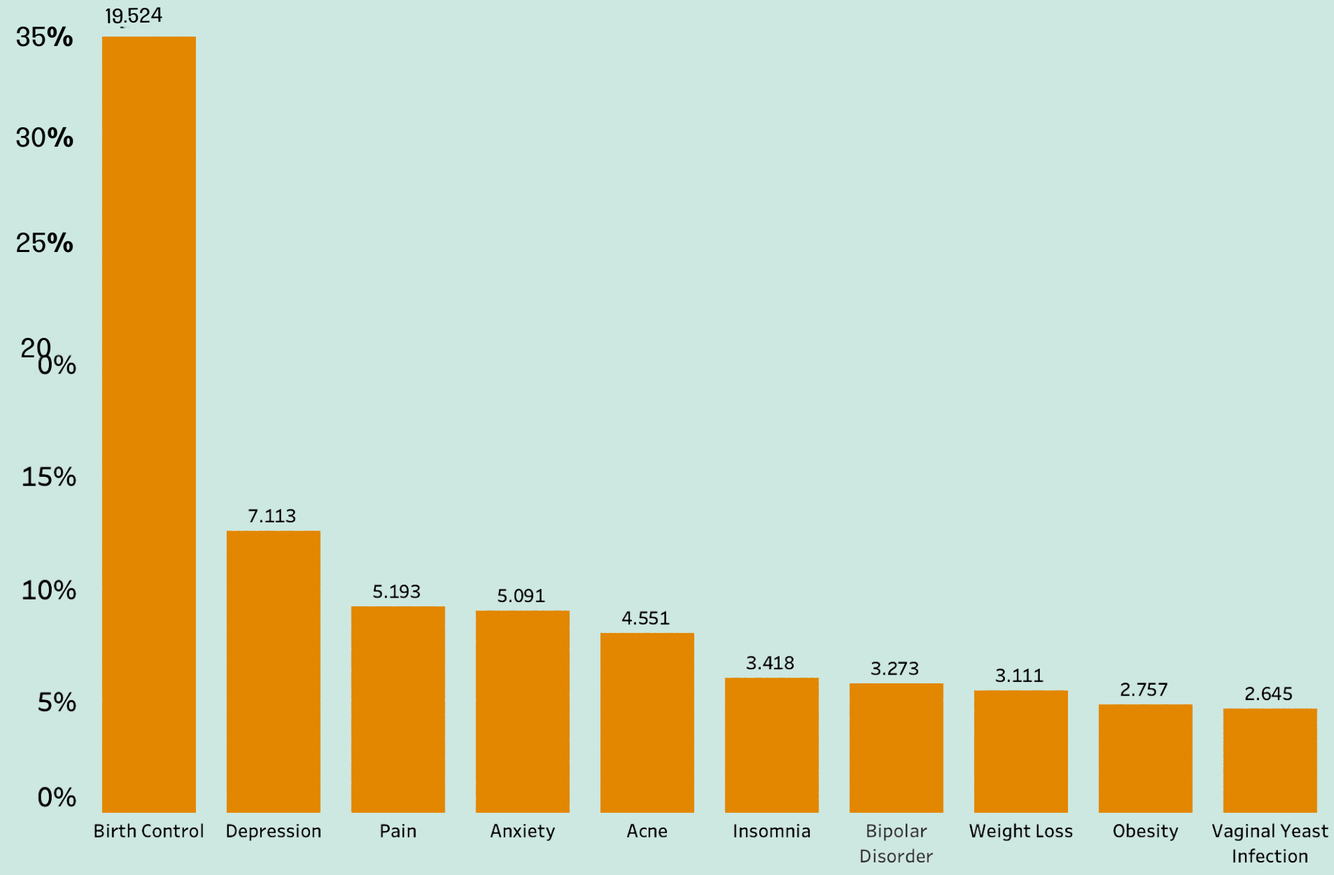
3,199
UNIQUE DRUGS

837
UNIQUE CONDITIONS

130,285
UNIQUE REVIEWS

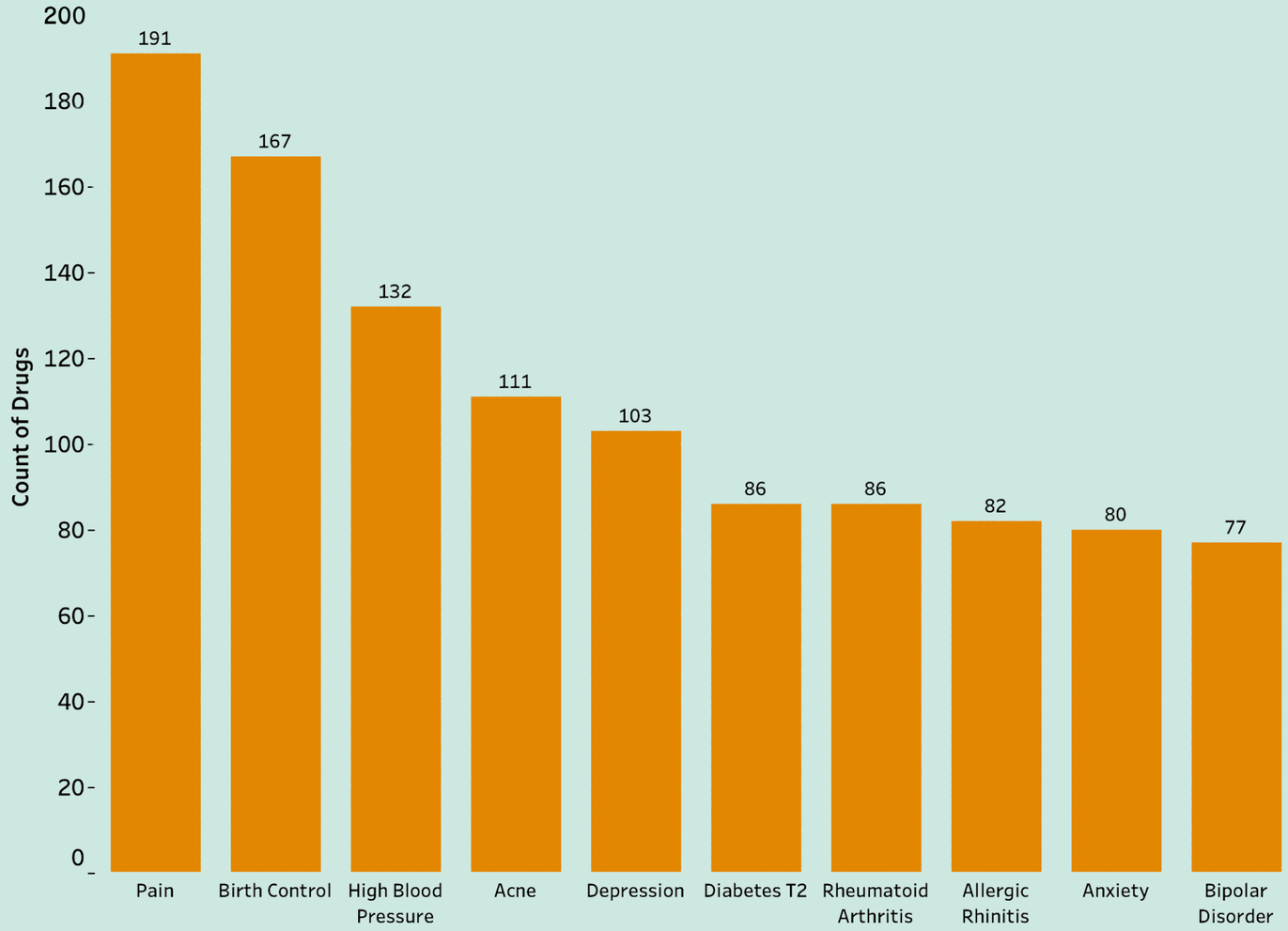
Drug types and medical conditions in the dataset

Top 10:
Review count per Condition



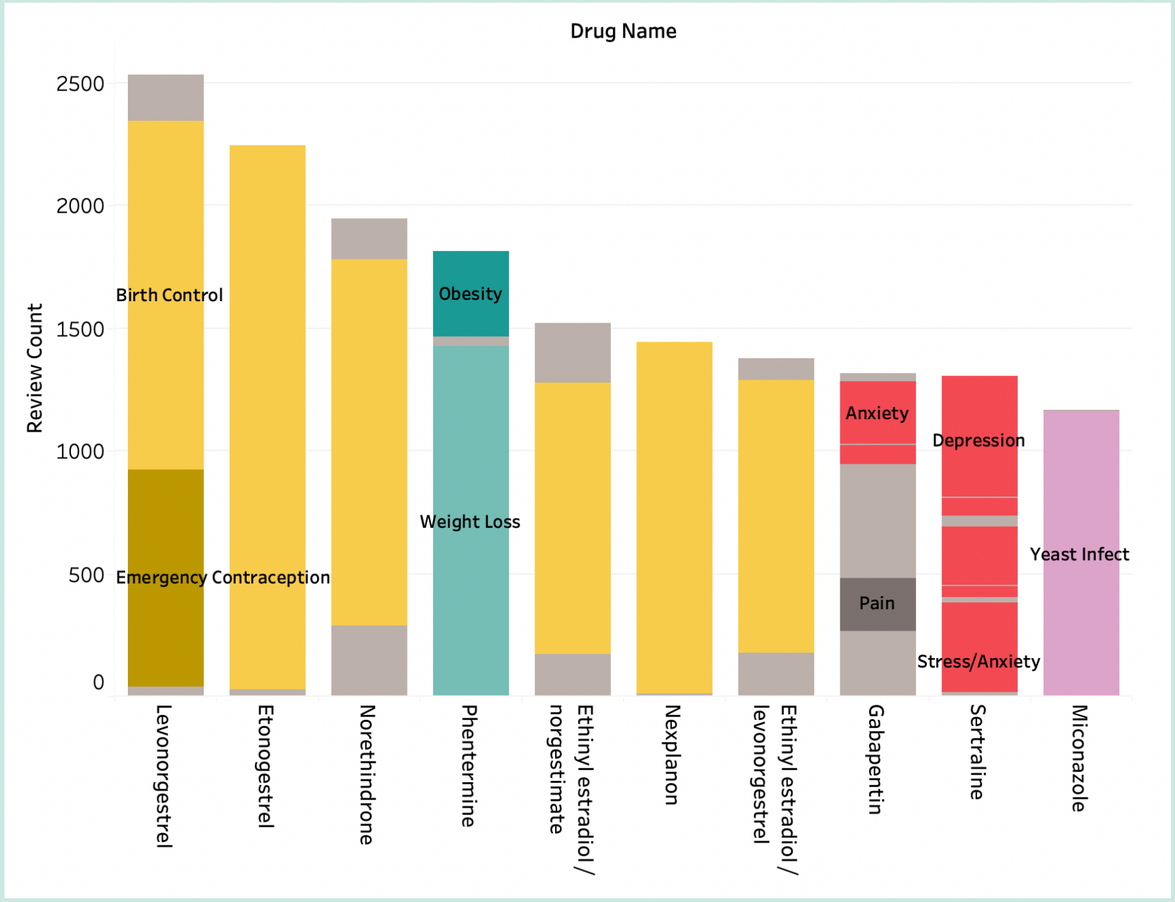
Drug types and medical conditions in the dataset

Top 10:
Drug count per Condition



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:

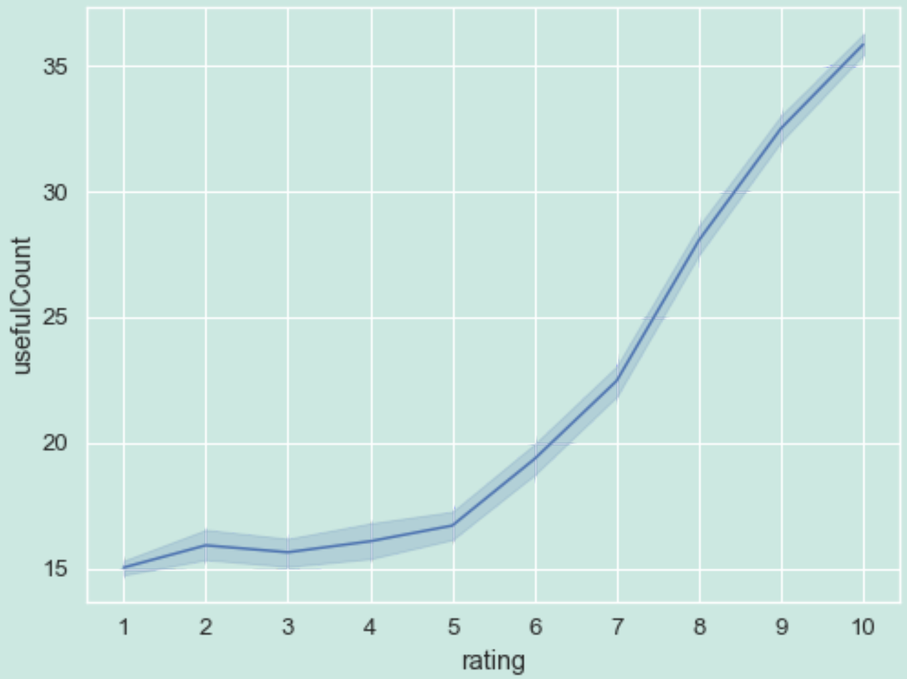
- Birth control and termination
- Mood disorders
- Weight loss

Drug recommender

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

Usefulness votes

Average of UsefulCount per Rating



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

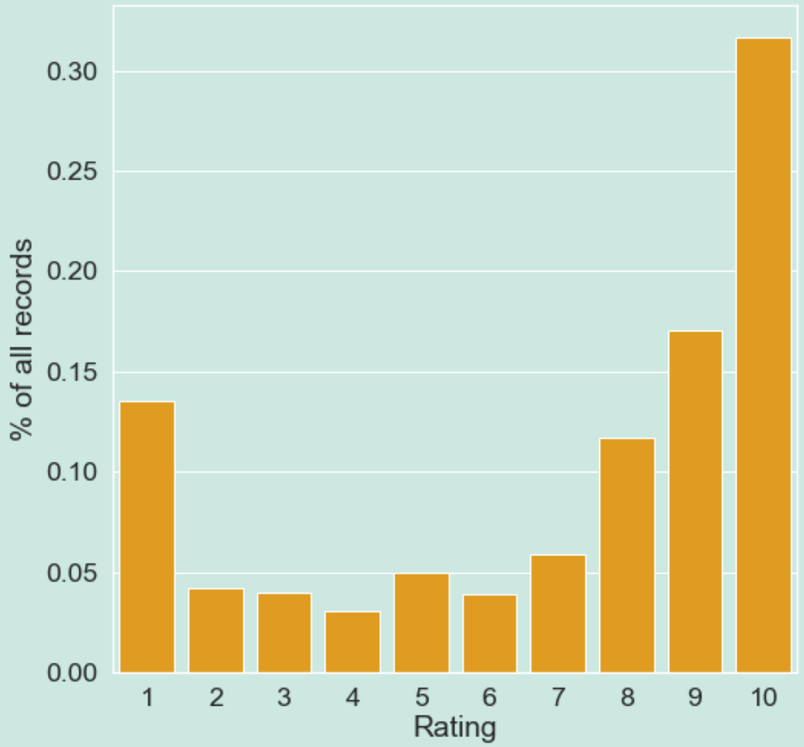
However, data is imbalanced towards positive reviews (see next slide):

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

=> No removing of records

★★★★★ Target variable: drug rating

Proportion of records per Rating



Feature engineering:

Rating of 1 - 6 ⟶ **"0"** class (negative/neutral)

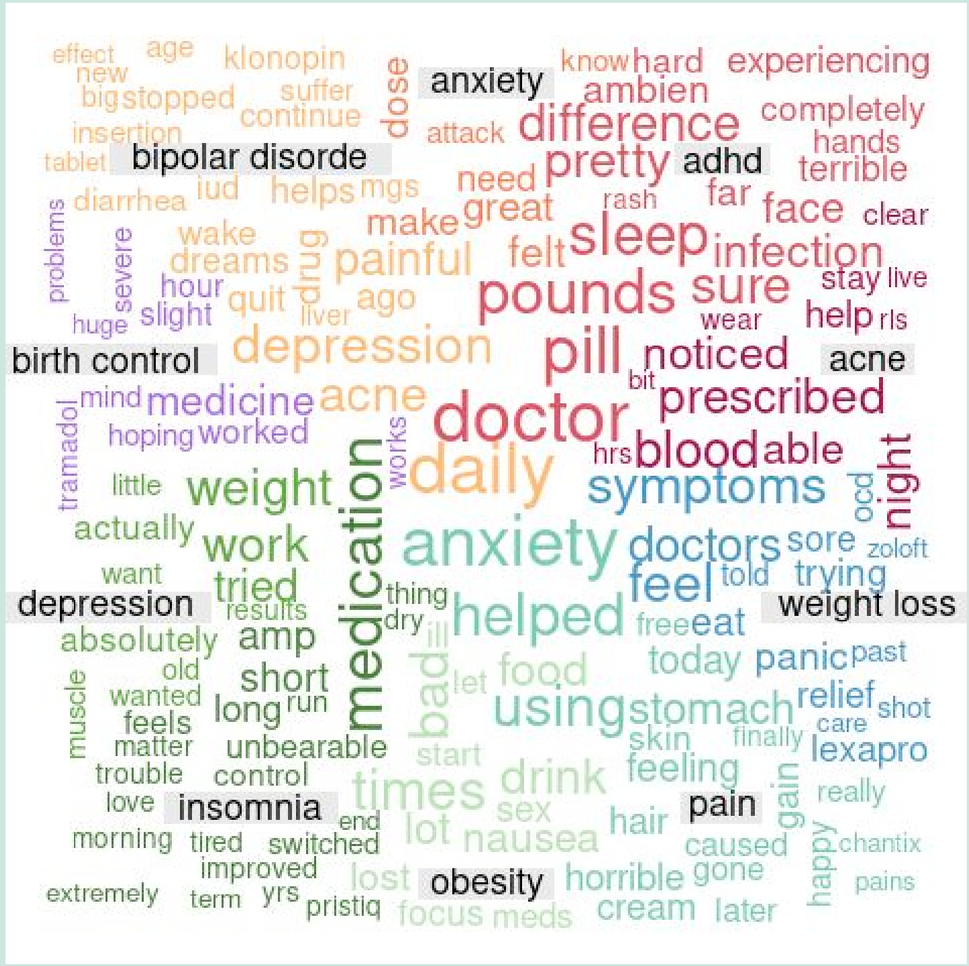
Rating of 7 - 10 ⟶ **"1"** class (positive)

Imbalanced data

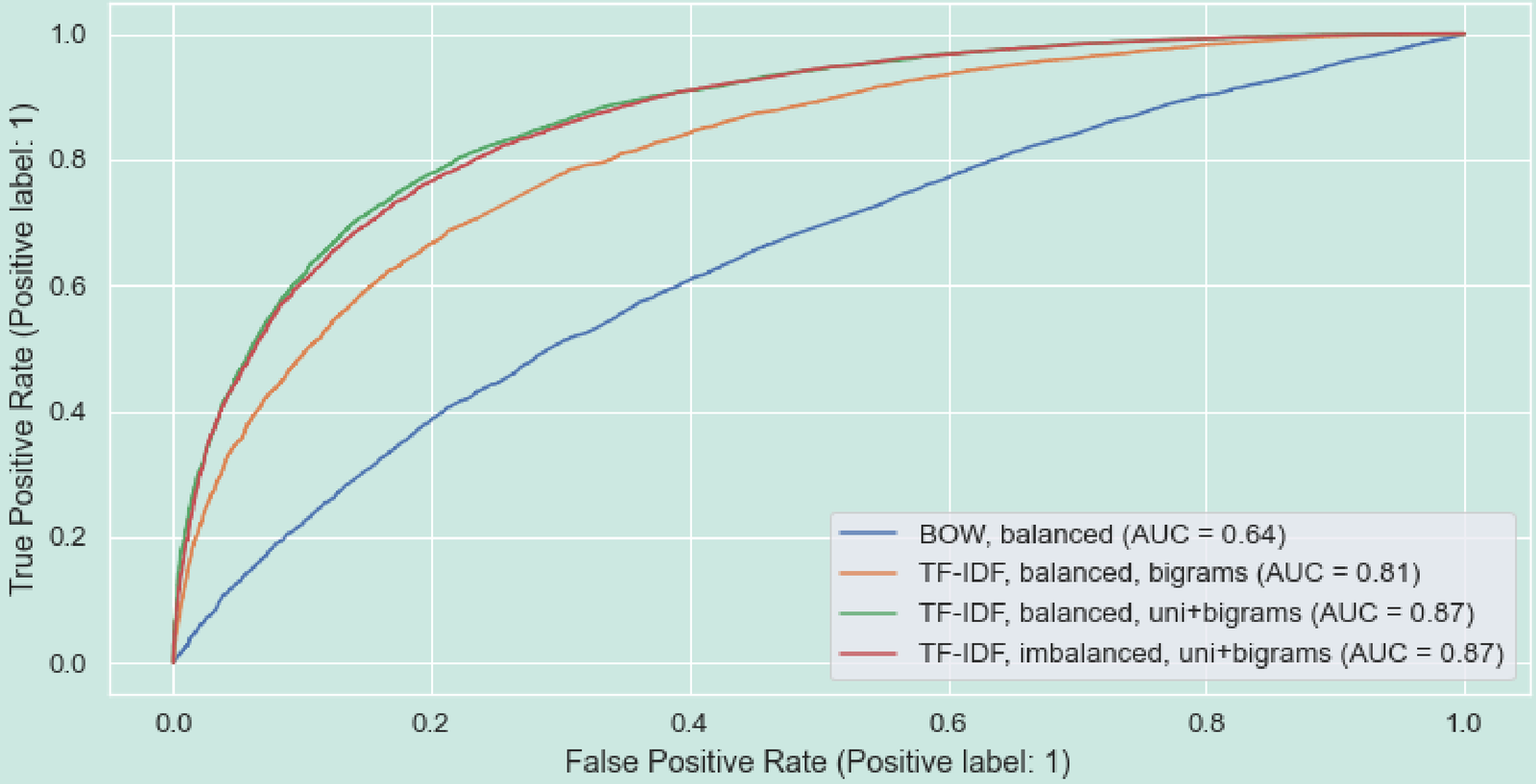
Class 0 still twice as big, so I ran the models on both a full and a downsample.

Word Clouds

Most frequent words associated with top 10 Conditions



Model Comparison



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

Limitations and Future Ideas

- General issues
  - Not filtering according to UsefulCounts
  - Not distinguishing a neutral category
  - All models are overfitting

- Modeling
  - Both Bag of Words and TF-IDF models lose a lot of semantic information
    - Word2Vec overcomes some issues by placing words in relation to one another, so give some idea of semantic proximity