How good machine learning algorithms are predicting the rating of drugs?

The story about a data miner's exploration in the NPL world

Lina Gao

Introduction

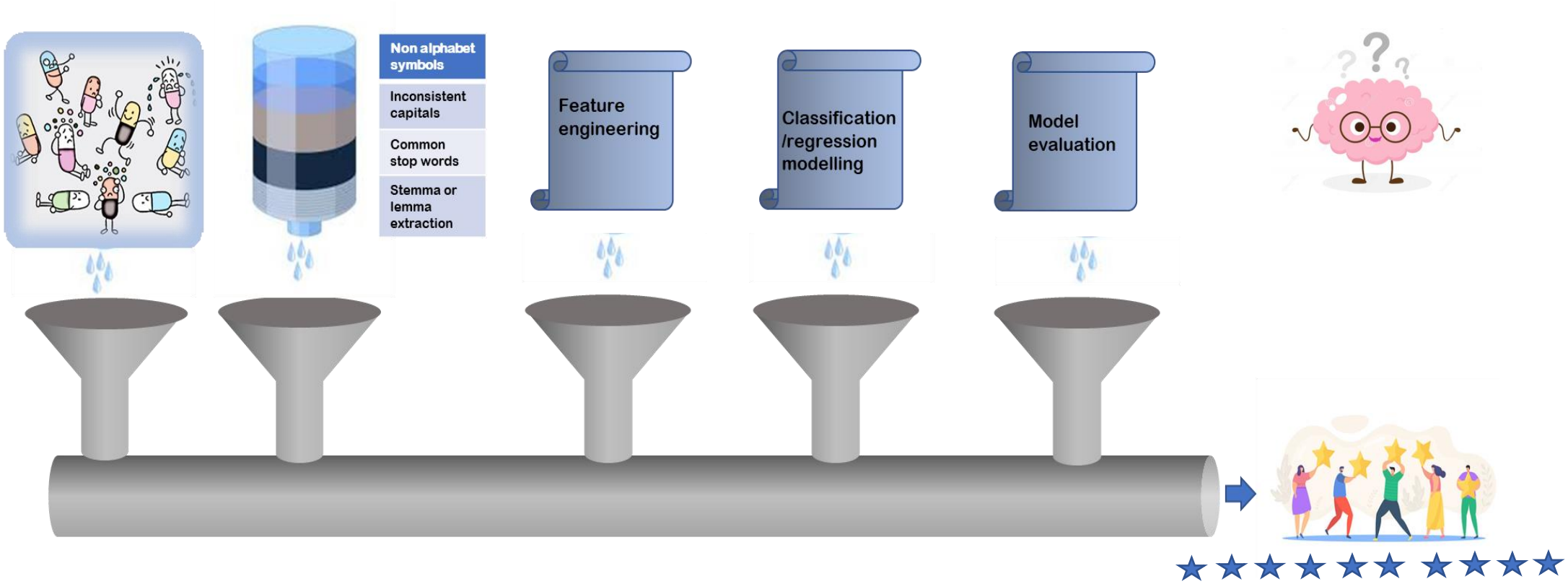
- Online user reviews in the pharmaceutical industry contain valuable real-life collective information about the drug's effectiveness and side effects.
- Because of the diversity and complexity of review data, precisely extracting sentiment information from them is still a challenge.
- Both lexicon + rule-based algorithms and machine learning algorithms have been applied in sentiment analysis of review data.

Motivation

- The primary question for this project is whether we can predict the rating of drugs using the review data by supervised regression/classification machine learning algorithms
- Although sentiment analysis has been widely applied in social media and eCommerce companies, its application in public health is still limited. Many stakeholders can be benefited from the sentiment analysis of drugs reviews:
    - the post-marketing analysis for pharmaceutical companies,
    - the clinical decision support system for health care providers,
    - the decision-making process for the patients.

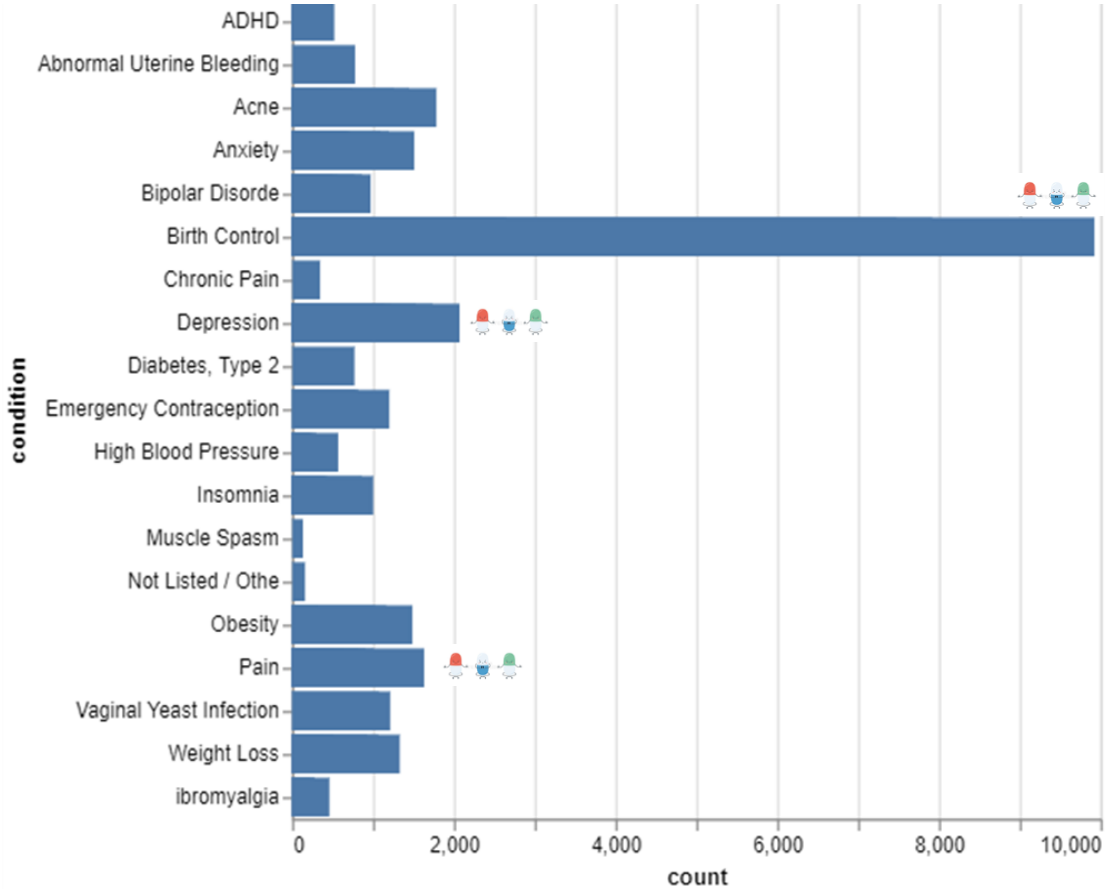
Task Overview

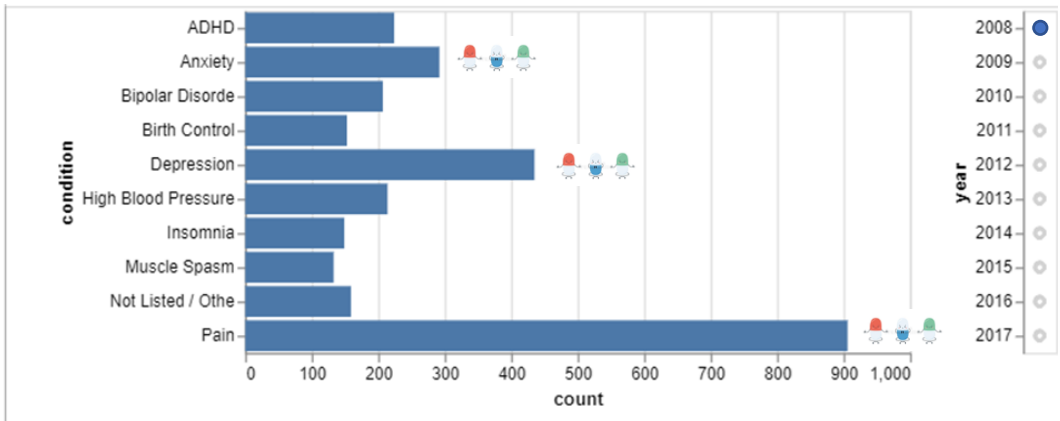
DIM: 213869 X 6

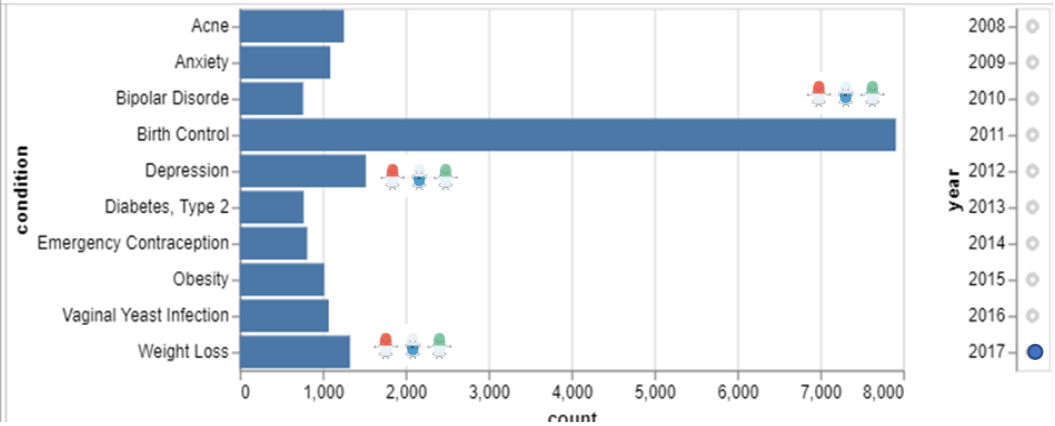


Q1. What does this data tell us?

Review N of the Top 10 conditions (916 in total)







- The most reviewed conditions had been greatly changed from 2007 to 2017.

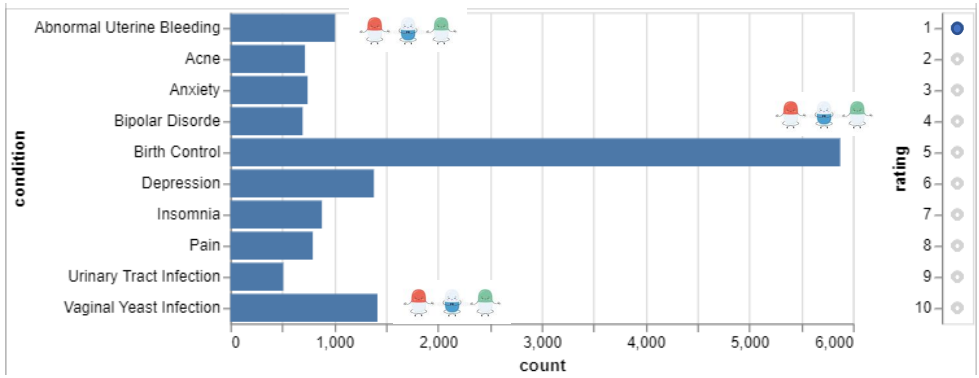
Drug Names & Conditions

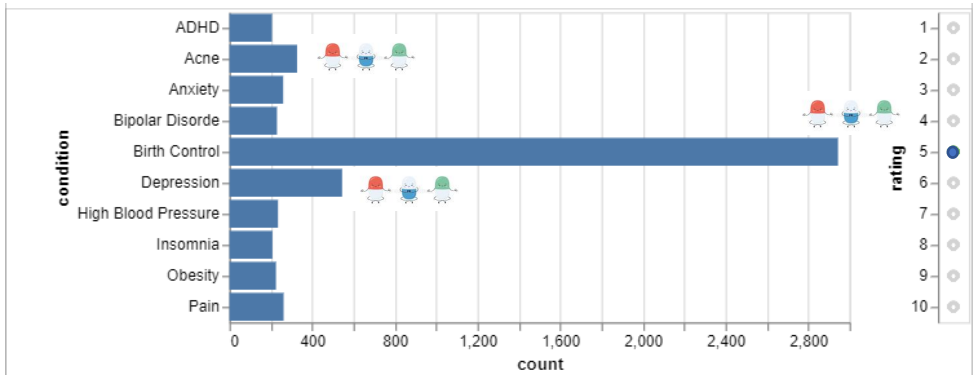
drugName	condition	review	rating
Bupropion	ADHD	72	72
	Anxiety	75	75
	Bipolar Disorde	41	41
	Depression	747	747
	Major Depressive Disorde	142	142
	Migraine Prevention	2	2
	Not Listed / Othe	5	5
	Obesity	6	6
	Panic Disorde	10	10
	Persistent Depressive Disorde	4	4
	Postural Orthostatic Tachycardia Syndrome	2	2
	Premenstrual Dysphoric Disorde	3	3
	Seasonal Affective Disorde	14	14
	Sexual Dysfunction, SSRI Induced	38	38
	Smoking Cessation	199	199

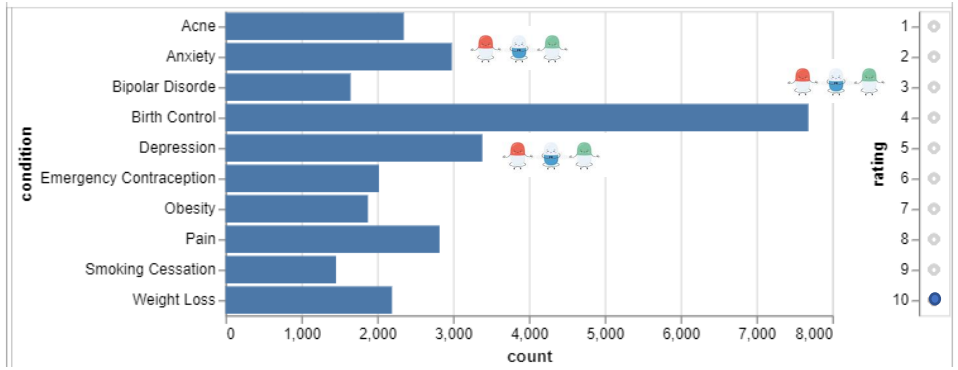
condition	drugName
ADHD	Adderall
	Adderall XR
	Adzenys XR-ODT
	Amantadine
	Amphetamine
	Amphetamine / dextroamphetamine
	Aptensio XR
	Armodafinil
	Atomoxetine
	Budeprion XL

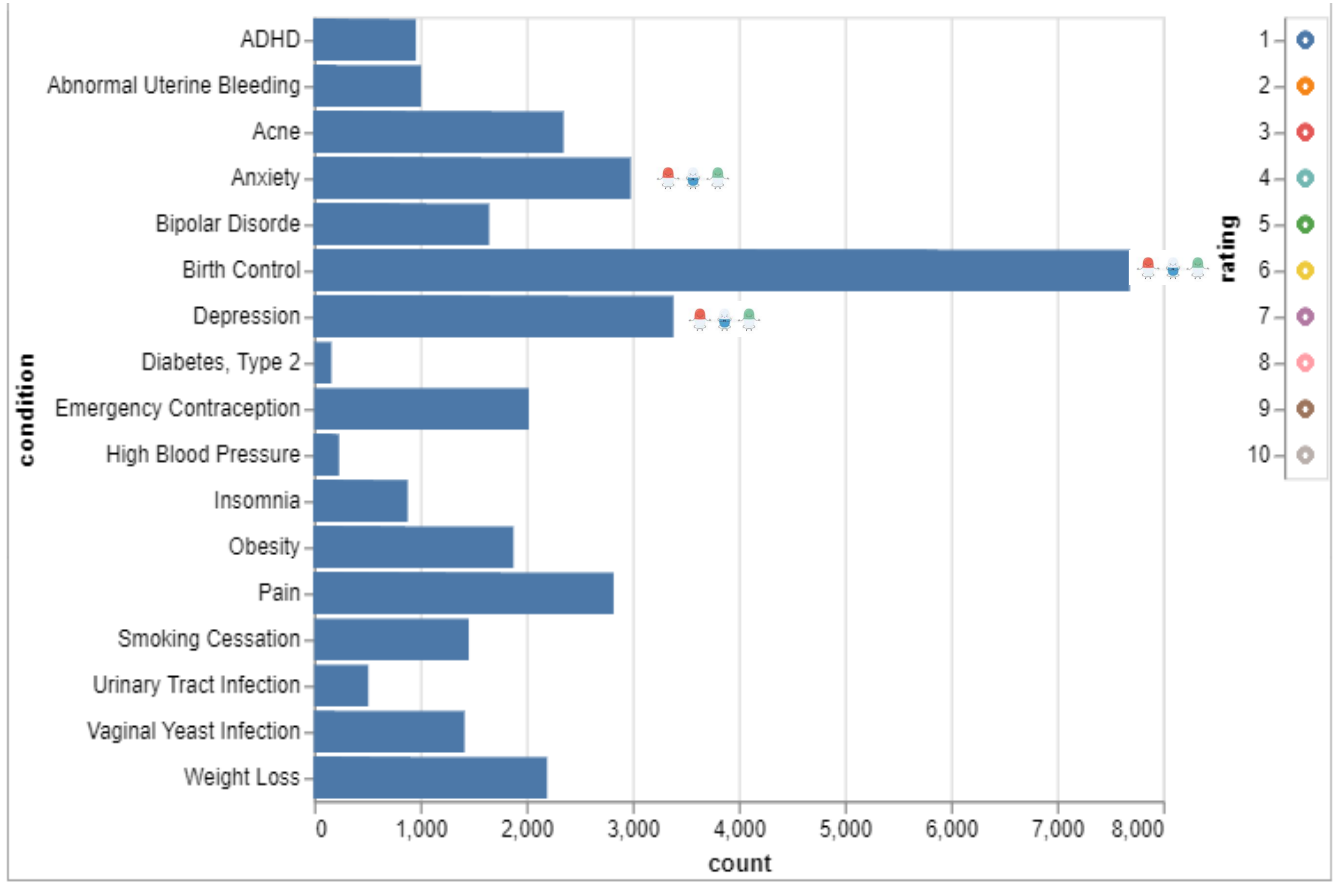
- One drug can be applied for multiple conditions. One condition can be treated by multiple drugs. There are totally 3667 drugs in this data set.

Ratings & top 10 Conditions



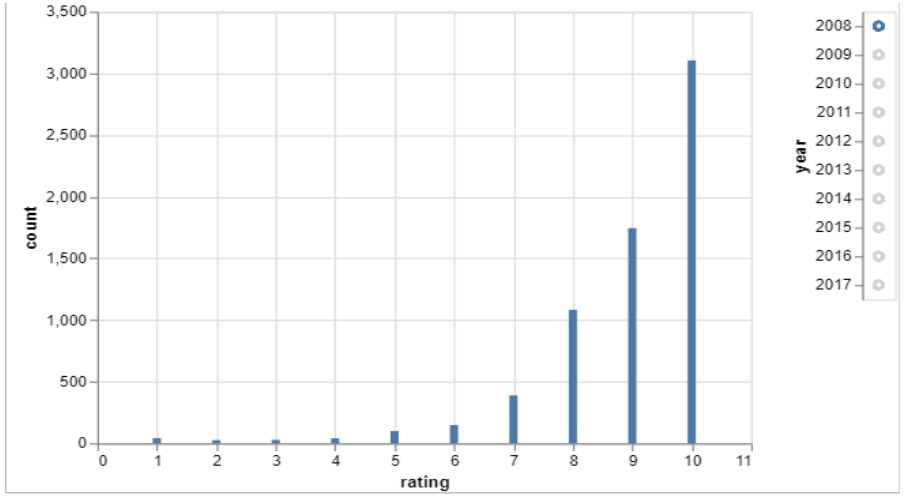


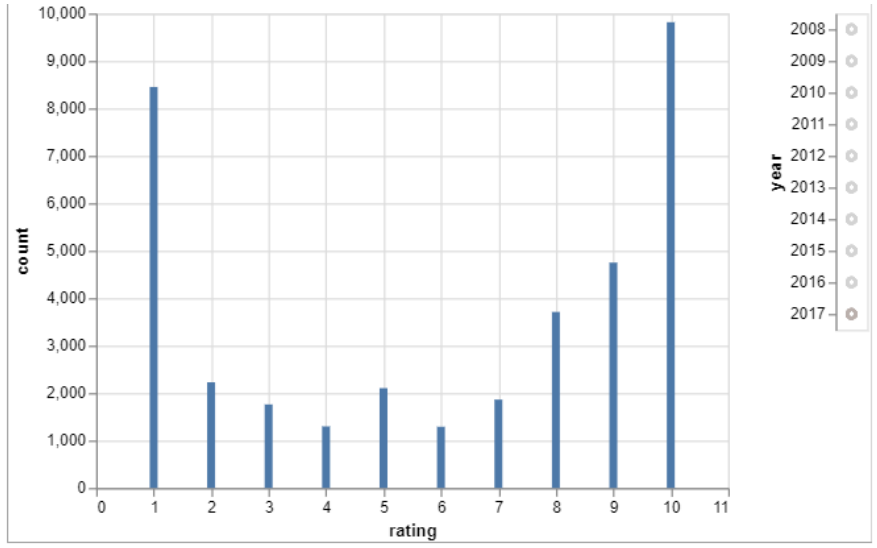


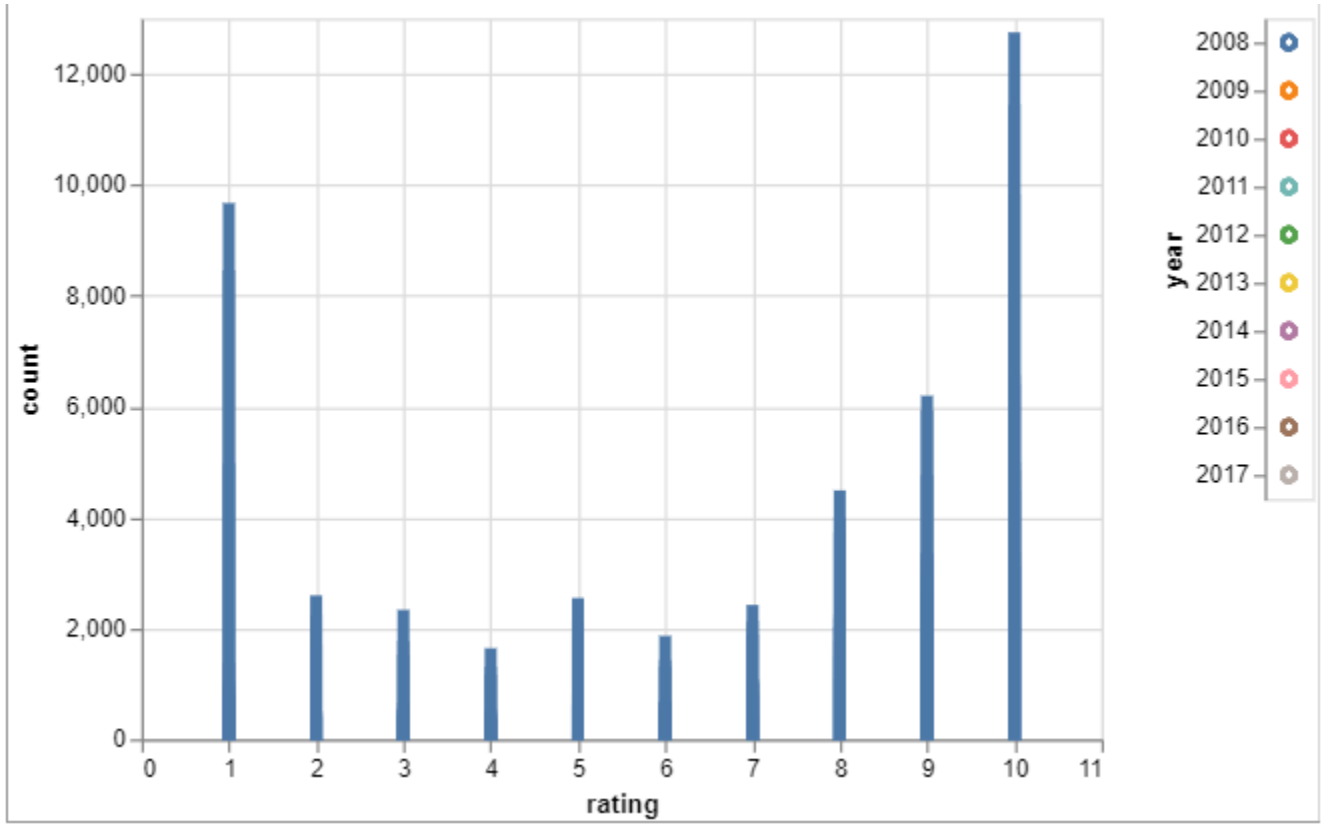


- The distribution of conditions across ratings is diverse.

Ratings & Years





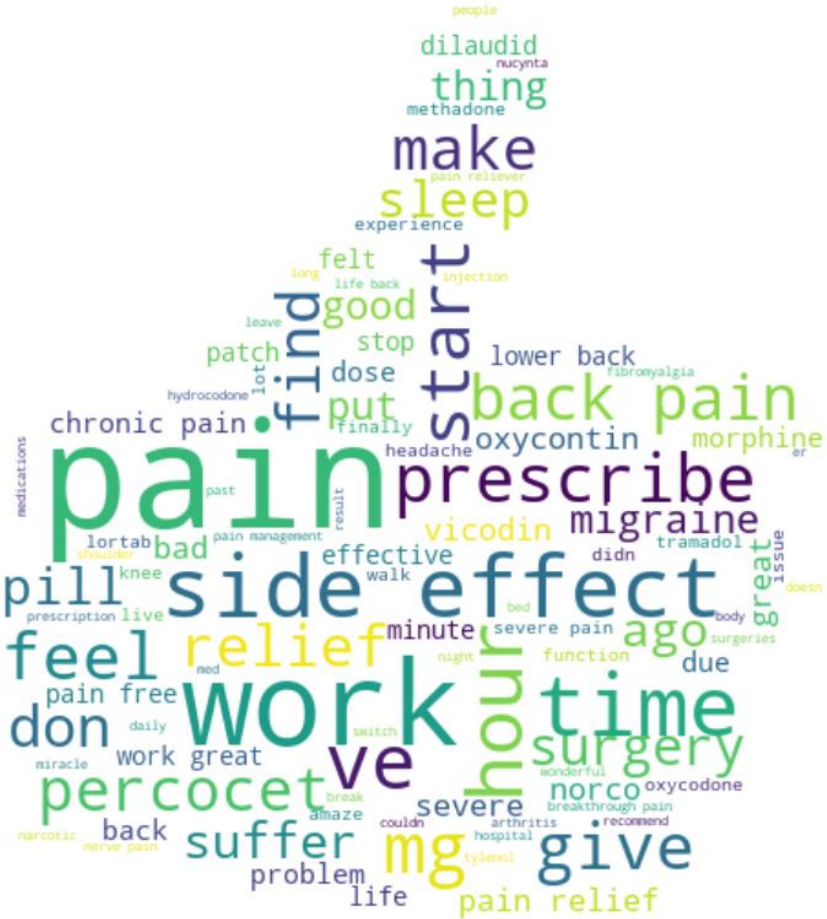


- Higher ratings of reviews are the dominant in 2007-2015. Polarized ratings turned into the primary contributor in 2016 and 2017.

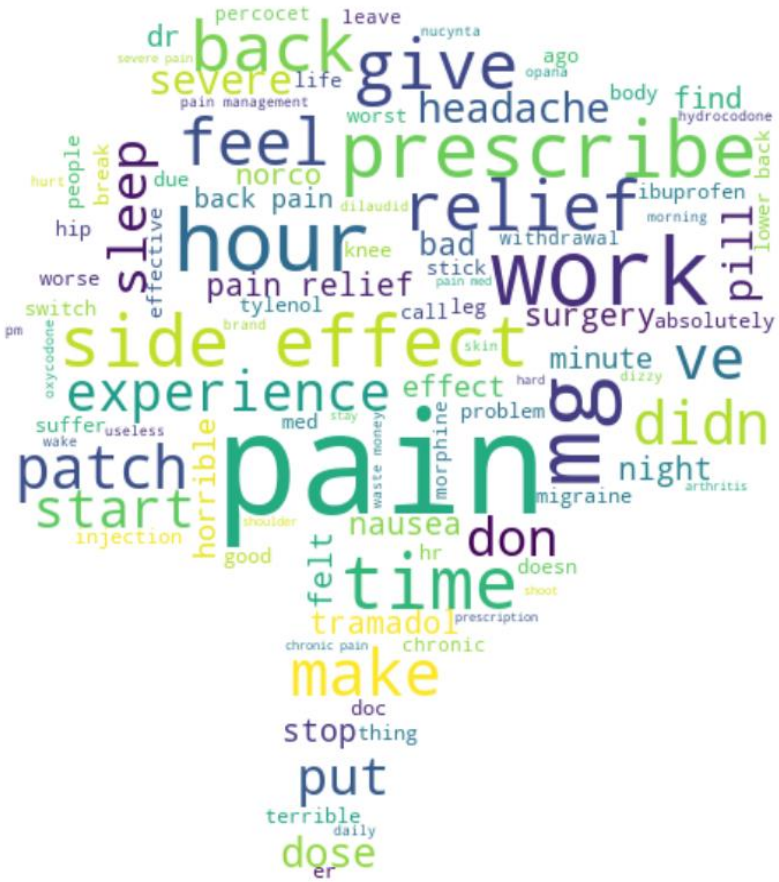
Reviews: Word Cloud for Pain Relievers as an Example (unigram)

- Without differentiating ratings, wordcloud plot can be a mixed bag of words. We can find both positive and negative words besides drug names and syndromes. After separating high rating reviews from low rating reviews, we can get more clear information.

Reviews: Word Cloud for Pain Relievers as an Example (unigram)

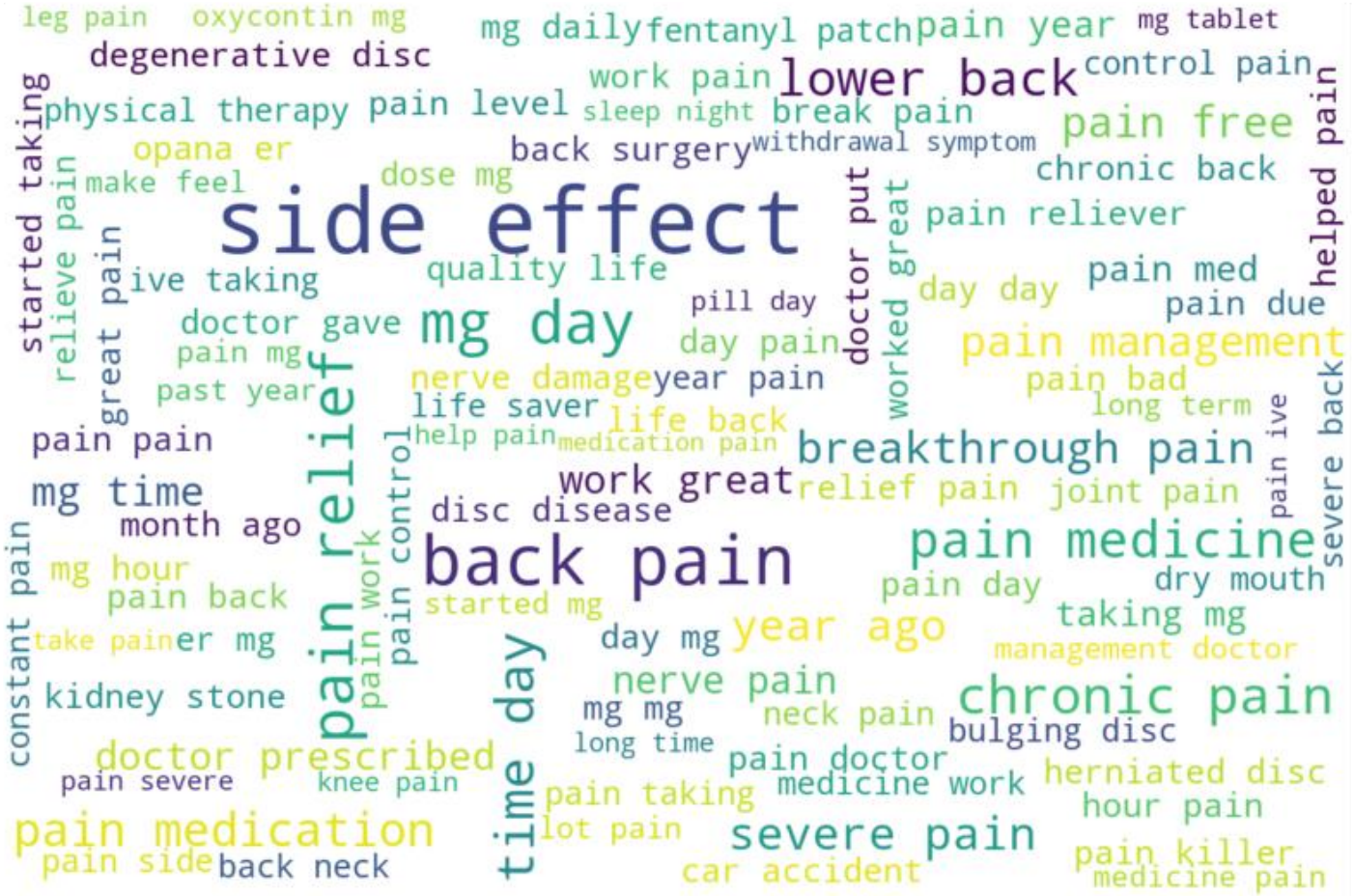


Ratings: 10-8



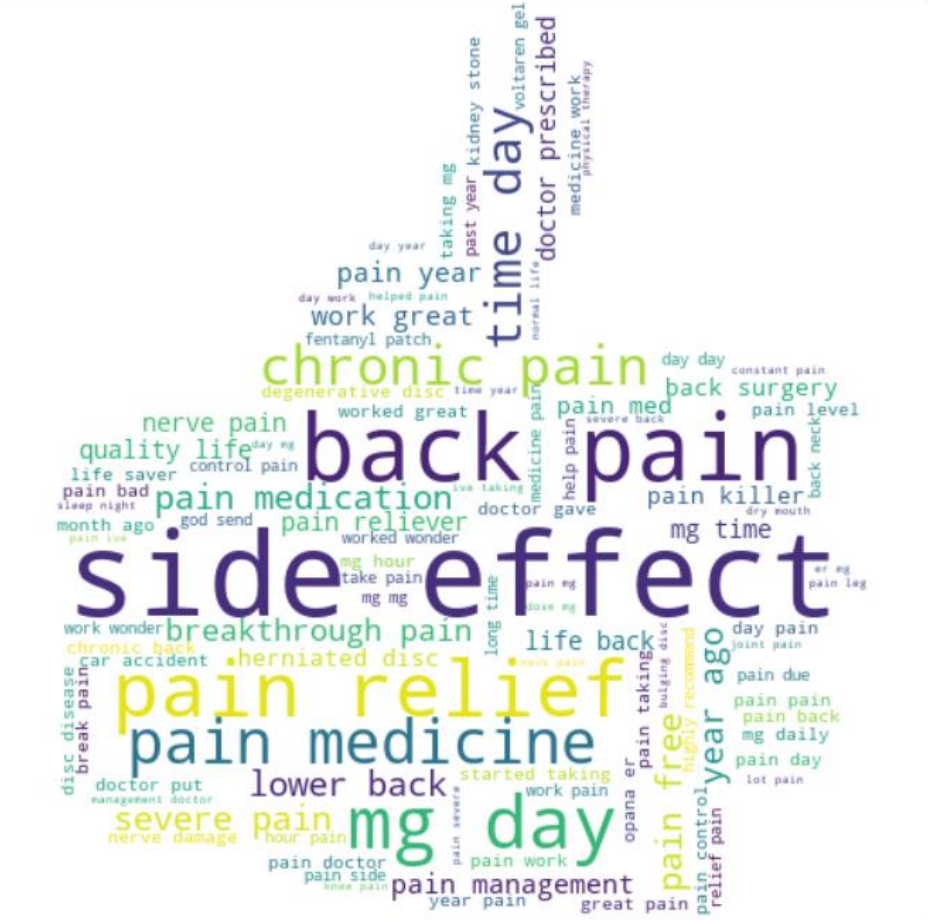
Ratings: 1-3

Reviews: Word Cloud for Pain Relievers as an Example (bigram)

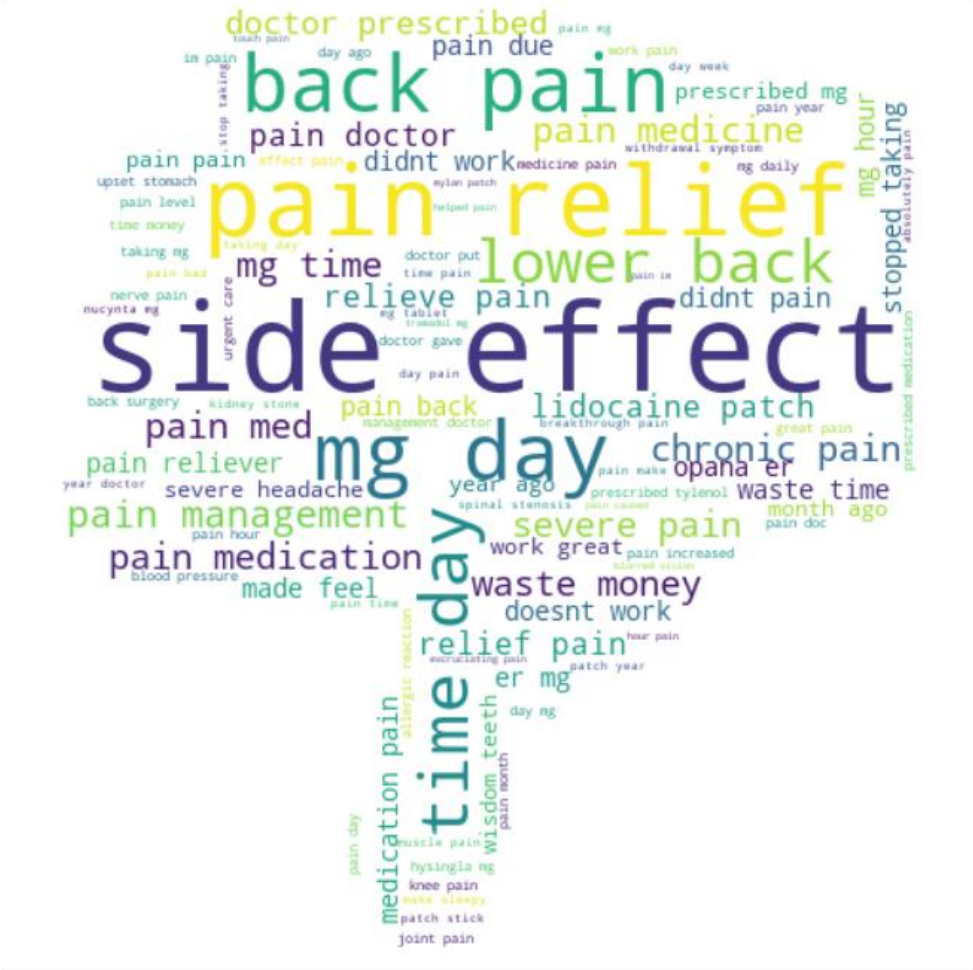


- We can get more clear information about drug effectiveness and side effect with bigram wordcloud than unigram wordcloud.

Reviews: Word Cloud for Pain Relievers as an Example (bigram)



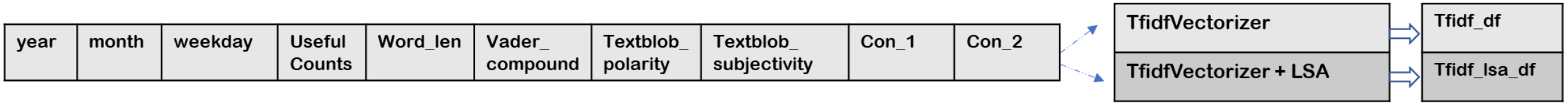
Ratings: 10-8



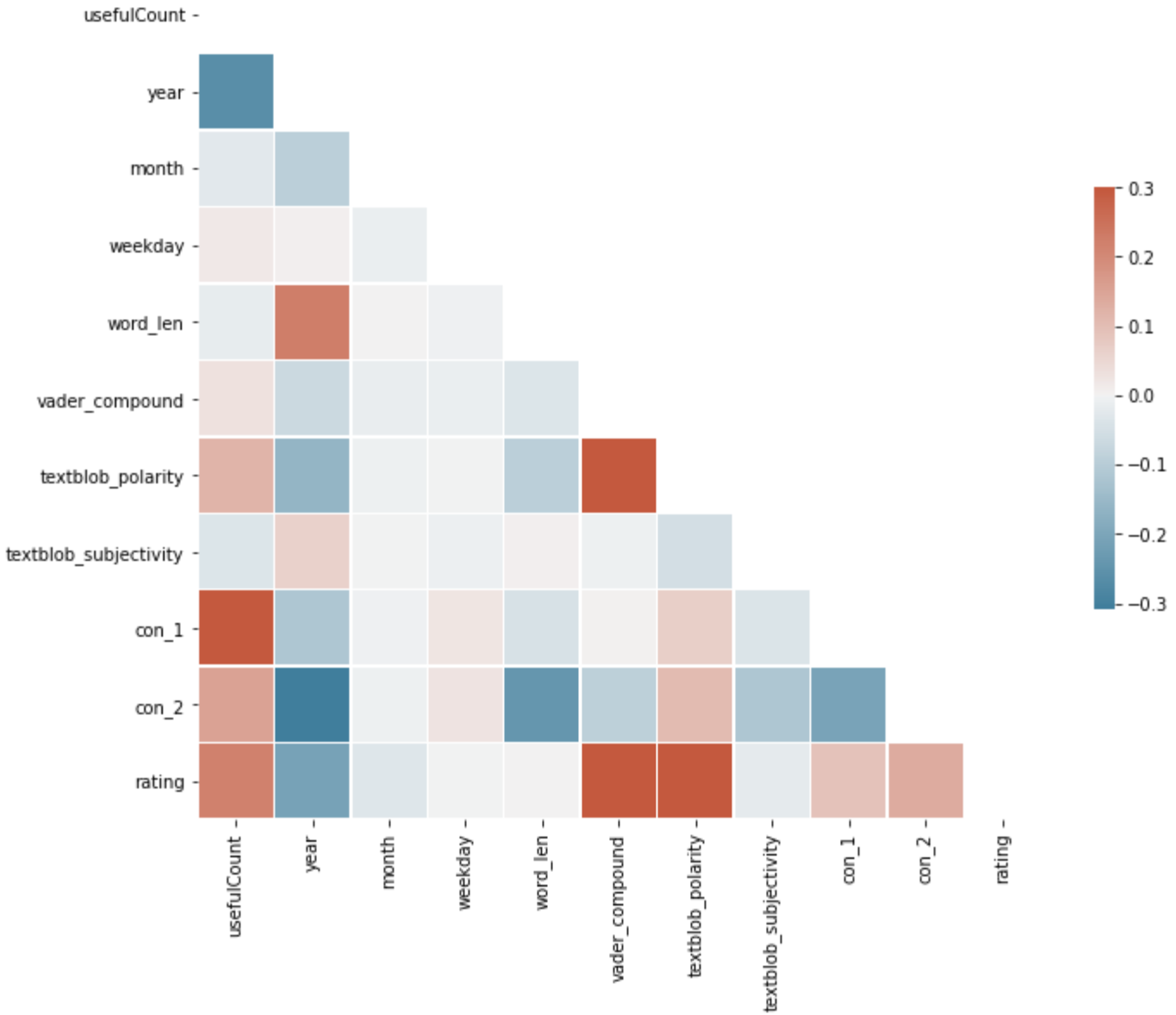
Ratings: 1-3

Q2: How can we extract useful features from the data for modeling?💡

- Choose the top 3 conditions data for modeling, birth control, pain and depression. Convert condition into dummy variables.
- Extract year, month and weekday from datetime.
- Extract word length of the reviews as features
- use VADER, to get the sentiment indicator as features
- use TextBlob, to extract sentiment information as features
- Use TfidfVectorizer transform review to get 288 new features. Combine these features with the original features to form tfidf_df
- Use Latent sematic analysis (LSA) to extract information from the review to get 100 new features. Combine these features with the original features to get tfidf_lsa_df

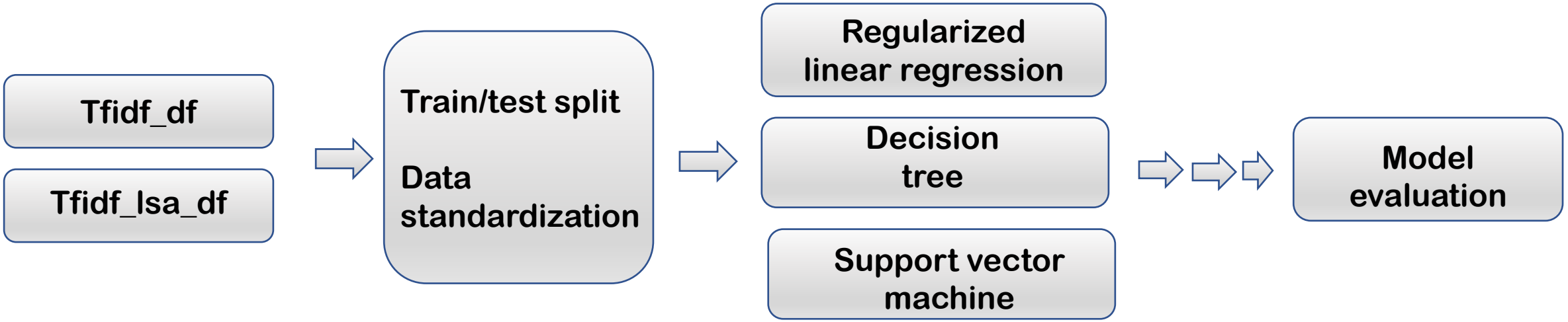


Feature Engineering: correlations between features



Q3: Should we develop regression model or classification model?💡

Regression Model

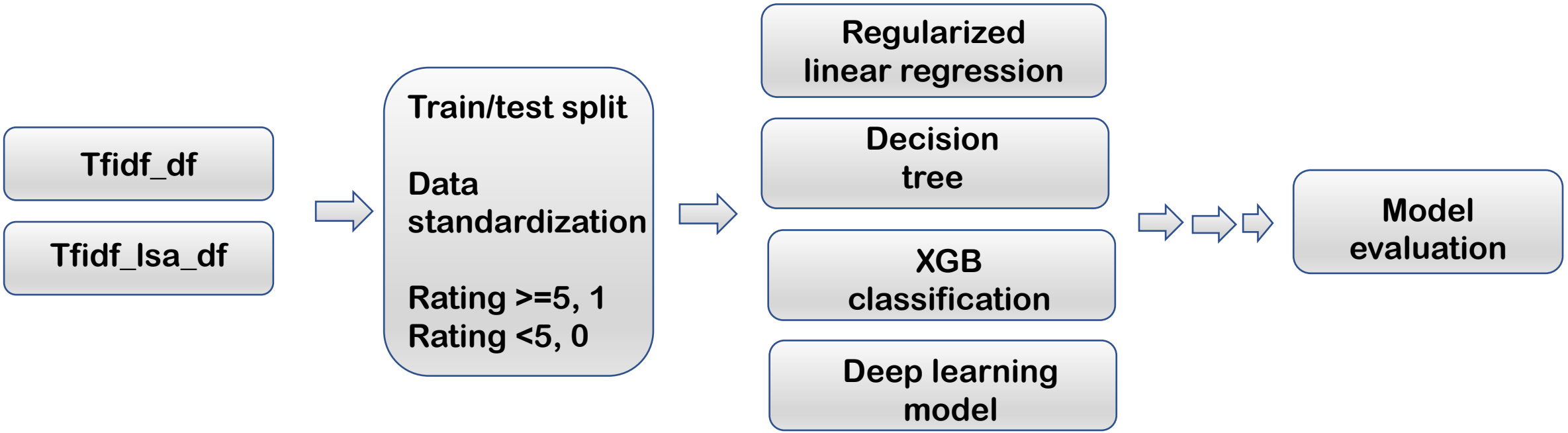


Regression model

Data	Algorithms	Run time	hyperparameter	MSE	MAE	R^2
Tfidf_df	Regularized linear regression	159	Alphas, L1_ratios	6.12	1.99	0.43
	Decision tree	97	Max_depth	5.03	0.97	0.53
	Support vector machine	1908	Default setting	6.12	1.99	0.43
tfidf_lsa_df	Regularized linear regression	51	Alphas, L1_ratios	6.38	2.04	0.40
	Decision tree	140	Max_depth	5.18	0.98	0.52
	Support vector machine	735	Default setting	6.38	2.04	0.40

Q3: Should we develop regression model or classification model?💡

Classification Model



Classification model

Data	Algorithms	Hyper-parameter	Accuracy	Precision	F1 score	Recall	AUC
Tfidf_df	Regularized logistic regression	C	0.809	0.837	0.869	0.903	0.867
	Decision tree	Criterion, Max_depth	0.788	0.82	0.855	0.894	0.817
	XGBoosting	Gamma, subsample, n_estimator, max_depth	0.919	0.934	0.943	0.953	0.953
	Deep learning	N/A	0.907	0.935	0.933	0.932	N/A
tfidf_lsa_df	Regularized logistic regression	C	0.806	0.832	0.867	0.905	0.856
	Decision tree	Criterion, Max_depth	0.809	0.842	0.868	0.896	0.818
	XGBoosting	Gamma, subsample, n_estimator, max_depth	0.926	0.933	0.948	0.964	0.971
	Deep learning	N/A	0.837	0.875	0.885	0.896	N/A

Q4. Conclusions?

- Classification model can give us decent results
- Features created by tfidf + LSA can produce comparable results as tfidf alone engineered features.
- XGBoosting is the best performer among all tested algorithms
- Deep learning models are the top models in this case.
- We don’t suggest to develop regression model considering the long run time and poor model performance.

Q5. Future work?💡

- Develop more comprehensive models including all conditions in the original data instead of top 3.
- Optimize the hyperparameters in deep learning models and test their performance again.
- Develop web application to give the predicted results directly to the stakeholders.
- Try auto training and fast deployment for state-of-the-art NLP models on Hugging Face