A deeper look into online reviews of mental disorders and their impact on Natural Language Processing

Christophe Friezas Gonçalves (c.friezasgoncalves@tilburguniversity.edu)
U-number: u346847
Group: 3

Abstract

Medical trials take up many resources to assess drugs. Given the accessibility of online reviews these resources can be reduced. To accurately assess the utility of these reviews, deeper understanding of the reviews impact on Machine learning algorithms is needed. Thus study focused on training two Machine learning classifiers on two different conditions, an unrestrictive data set of 5 mental disorders online reviews and the same data set run through excessive preprocessing. The results show a higher accuracy on the unrestrictive data with unigrams, the down side being meagre recall and understanding. The optimal understanding lies in a restrictive data set with the use of bigrams. Ending in the conclusion, that higher word combinations bring meaning with the drawback of generalizability.

Keywords: Online reviews; Mental disorder; TF-IDF vectorization; Pre-processing

Introduction

Mental health is a topic that affects each and every human. Given the recent years, global mental health was challenged in many ways, yet there are people that struggle each and every day with their mental health. People suffering from ADHD, Depression or Schizophrenia for example need special medication to live a close to normal live. New medicine gets developed at a fast pace yet it has to overcome testing and clinical trials to see if it is fit for distribution. This does not mean that these medications come without discomfort or side-effects, thus online reviews can help to improve the understanding and longevity of these drugs(Adusumalli et al, 2015). Online reviews give access to large amounts of information on patients with mental disorder, yet users may forget to score their overall satisfaction. This may lead to faulty or misleading results when it comes to overall evaluation or usage of the datasets.
Some research has been done on online reviews as well as the data set used in this study(Ligthart et al, 2021; Garg, 2021; Vijayaraghavan & Basu, 2020). Given the different use of language in the reviews conventional Natural Language processing in medical environments have to be adjusted (Denecke & Deng, 2015). Nonetheless, this study is meant to deepen the understanding of what impacts the Machine learning algorithms and help improve the categorisation of the online reviews.

Research Question

Given specific mental disorders (ADHD, Anxiety, Bipolar disorder, Depression, and Schizophrenia), what are the main textual features that can improve the understanding of user satisfaction for certain drugs?

Methods

Data set

This study uses the drugs.com data set, as provided at https://github.com/fzamberlan/scrapdrugs. The dataset consists of textual reviews, reflecting the experience and satisfaction of the user with the administered drug, as well as scores from 0 to 10, the name of the drug, the period of consumption, the date of entry, a numerical value showing how many people found the review helpful. The dataset will be divided into 5 datasets for each disorder(ADHD, Depression, Schizophrenia, Anxiety, Bipolar) with the 10 most reviewed drugs per disorder.
The scores were divided into three categories, positive(2) for values ranging from ten to seven, neutral(1) ranging from seven to five and negative(0) ranging from four to one. The zero score reviews were omitted due to mislabeling by the users.
The textual reviews were stripped of meaningless words as a first step. After continuous evaluation of the models, words were omitted that all three categories(Positive, Neutral, Negative) had in common for strong predictors resulting in the restrictive condition. The reviews underwent TF-IDF vectorization after the cleaning process to be in a numerical format for the Machine learning algorithms and split into train and test sets(Garg, 2021).

Table 1: Datasets

Disease	Overall data	Train set	Test set
ADHD	1944	1555	389
Anxiety	3896	3116	780
Bipolar	2294	1835	459
Depression	2656	2124	532
Schizophrenia	387	309	78

Bias Prevention

Given the oversaturation of positive reviews, the models developed a bias towards classifying each entry as positive. The models had high accuracy on the test and training set(>.70), coming from the fact that over 70% of the test set

had a positive label. The confusion matrix showed the models to be only predicting positive outcomes.

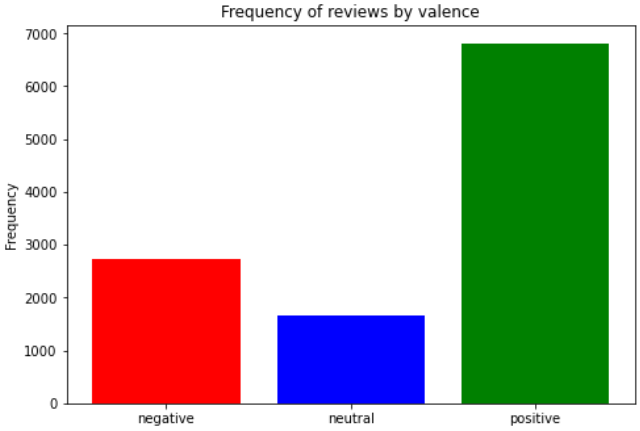


Figure 1: Overall Positive, neutral and negative Sentiment.

To prevent this bias a random over sampler was applied to even out the negative and neutral reviews to the positive ones. This lowered overall scores of the models, but changed the confusion matrix to show a balance in predictions.

Models

The 5 train sets were then used to train a Random Forest and a Multinomial Naïve Bayes Classifier. The models are a means to comparison between different conditions which is why the hyperparameter tuning was left to a minimal. A grid search 5 fold cross validation gave us the hyperparameter for max_depth(100) and n_estimators(500) for the Random Forest Classifier on the Schizophrenia train set. The Multinomial Naïve Bayes Classifier was implemented with default hyperparameters.

The 5 models were trained on unrestrictive data, meaning minimal word removal to encourage the creation of meaningful trigrams and bigrams. The models were then trained on the restrictive data. Both conditions were run with unigrams, bigrams and trigrams and evaluated on accuracy precision and recall scores.

Feature extraction and selection

From these trained models, the 10 most important predictors were extracted and plotted. For the Random Forest the overall predictors were extracted and for the Naïve Bayes the 10 most important features were extracted and plotted per sentiment category.

Results

Given the two conditions, we evaluate their accuracy based on contrast(Table 2. & 3.). The unrestrictive scores are overall higher. Depression bigrams being the only outlier. Unigrams performed better than bigrams and trigrams, with the latter one performing the worst. Highest accuracy resulted on the Anxiety and Bipolar Unigram models on the unrestrictive data (accuracy avg. = 0.67).

Table 2: Random Forest Unrestricted – Restricted.

Test accuracy	Unigram	Bigram	Trigram
ADHD	0.062	0.093	0.039
Anxiety	0.073	0.038	0.409
Bipolar	0.098	0.002	0.221
Depression	0.007	-0.026	0.115
Schizophrenia	0.071	0.007	0.125

Table 3: Naïve Bayes Unrestricted – Restricted.

Test accuracy	Unigram	Bigram	Trigram
ADHD	0.038	0.131	0.029
Anxiety	0.03	0.294	0.16
Bipolar	0.063	0.23	0.107
Depression	0.023	0.208	0.118
Schizophrenia	0.032	0.085	0.008

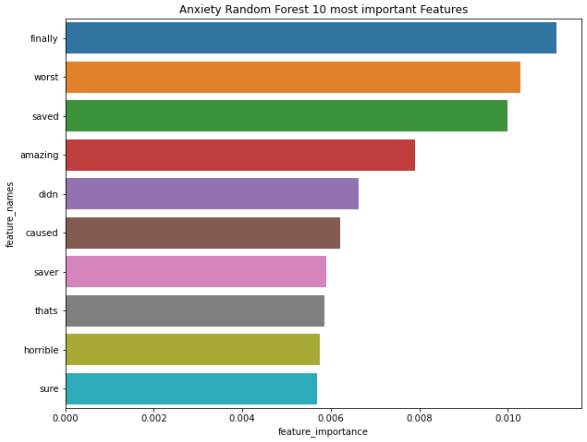
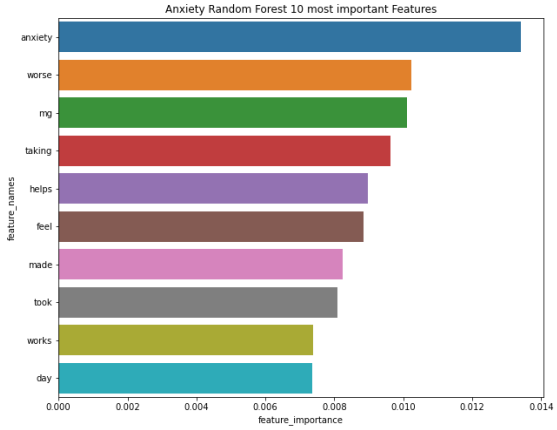
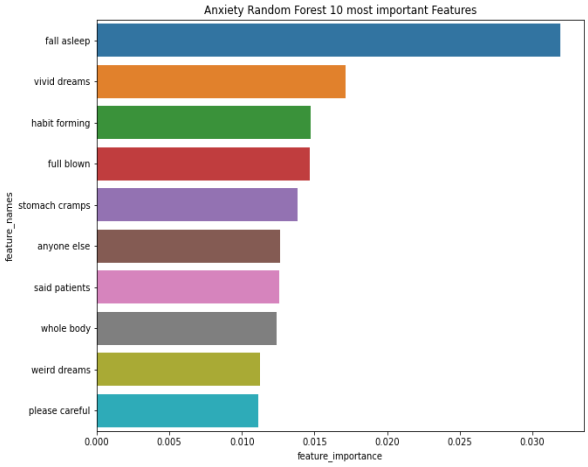
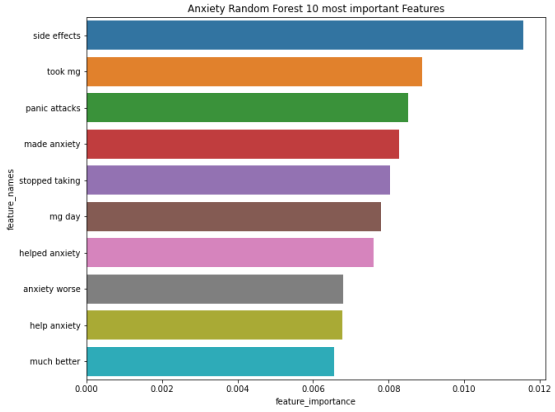
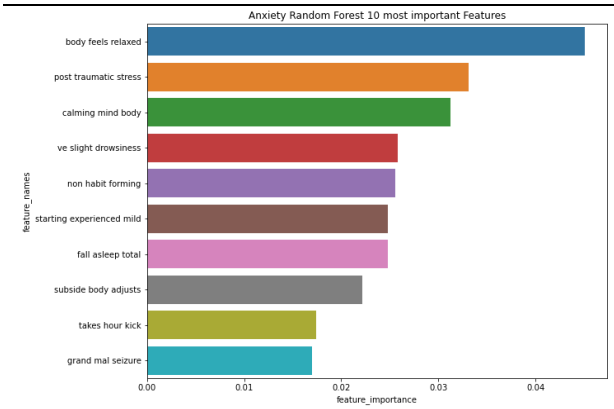
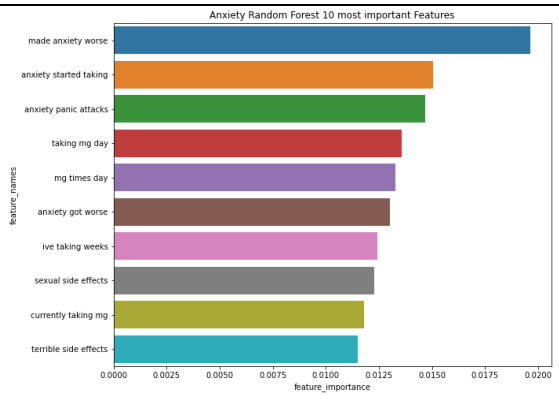
The feature extraction resulted in many combinations of words for trigrams and bigrams. For the unrestrictive data set, these combinations ranged from side effects to middle parts of the sentence, i.e. “taking mg day”. The unrestrictive models had many words in common as main predictors for the three sentiment categories. The Anxiety feature graph shows one example of the 10 most important words across the different conditions(Table 4.).

Discussion

Given the results obtained by the study, what are the main textual features, that can improve the understanding of user satisfaction for certain drugs? The results show a slight increase in accuracy for the unrestrictive data, which would lead one to believe, that the unrestrictive data was a better input to our models, yet after a deeper look into the confusion matrices it was noticeable that the models developed a bias towards predicting one category. Neutral had a low recall (>0.1) for unigrams on the unrestrictive data, however, it had an increased recall on the restrictive data. Inversely bigrams and trigrams showed the same pattern on the opposite data sets.

The restrictive Trigrams and bigrams show meaningful word combinations which are also present in the unrestrictive data, yet, to a different extent, there are more combinations bringing no interpretable results, for example “I ve taking weeks”. The textual features give less meaning and understanding of sentiment when implemented as

Table 3: Anxiety feature importance graph

Excessive word removal	Simple word removal
	
	
	

unigrams. Words like “horrible”, “worst” and “finally” are high predictors yet do not convey a lot of meaning on the overall treatment of the patient. Bigrams and trigrams convey that missing understanding, they talk a lot about side effects and positive outcomes, “trouble sleeping” and “killed sex drive” as well as “within minutes”. These combinations bring understanding to the researcher and insight into the resulting sentiment.

Conclusion

User satisfaction with medical treatment has many facets, that can mainly be understood through a restrictive dataset and by the use of unigrams and bigrams depending on the goal of the researcher/doctor. Bigrams incorporate the use of negation and exaggerations into the features, helping to differentiate good effects from bad ones for the optimal resulting interpretability of the features.

The study shows the impact of word combinations in different data sets of mental disorders. Showing various results across the restrictive and unrestrictive data sets and the three word combinations(Unigram, bigram, trigram). The resulting graphs give a wide array of different results showing the main trend that understanding lies in a polished data set and a higher word combination, yet the word combination complexity lies in a trade off with generalizability.

Limitations and further research

A limitation of this study was the use of the data set. Using only the 10 most reviewed medications per Disease, the Schizophrenia data set ended up small in size, hence the models were trained on a meagre train set affecting the accuracy. A bigger data set may influence the resulting textual features, which would be interesting to explore. A second limitation would be the missing Lemmatizer and/or Stemmer, pre-processing methods used in Natural Language Processing. The study omitted both after testing and getting confusing results, to guarantee interpretable results. Further studies may be conducted on looking deeper into the mention of side effects or medical terms and their impact on model evaluation.

References

Adusumalli, S., Lee, H., Hoi, Q., Koo, S., Tan, I., Ng, P. (2015). Assessment of Web-Based Consumer Reviews as a Resource for Drug Performance, *Journal of Medical Internet Research*, 17(8):e211 DOI: 10.2196/jmir.4396

DOI:10.1109/Confluence51648.2021.9377188

Denecke, K., Deng, Y. (2015). Sentiment analysis in medical settings: New

opportunities and challenges. Artificial Intelligence in Medicine, 64(1), 17–27.

https://doi.org/10.1016/j.artmed.2015.03.006

Garg, S. (2021). Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning. *11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*

DOI:10.1109/Confluence51648.2021.9377188

Ligthart, A., Catal, C., Tekinerdogan, B. (2021). Systematic reviews in sentiment analysis: a tertiary study. *Artificial Intelligence Review*. 54. 1-57.

https://doi.org/10.1007/s10462-021-09973-3

Vijayaraghavan, S., Basu, D. (2020). Sentiment Analysis in Drug Reviews using Supervised Machine Learning Algorithms. *ArXiv, abs/2003.11643*.

https://arxiv.org/pdf/2003.11643.pdf