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?

Ann-Kristin Balve (u649435, group 3)

Abstract

Abstract Analyzing psychic drug reviews using sentiment analysis can help understand textual predictors for patient satisfaction. In this study, drug reviews were used to train a Naïve Bayes and Random Forest classifier and visualize the most important features to better understand the internal workings of the algorithm. Different feature sets, consisting of various stopword and n-gram combinations, were compared. The results show that the choice of the textual features affected the model performance with best results for unigram and unigram-bigram combinations. For some diseases, more restricted vocabulary sets furthermore improved the model predictions. Visualizing the textual features gave useful insight into predictors of drug satisfaction, revealing the potential of sentiment analysis to analyse determinants of drug success which can be used to provide mental disorder patients with improved treatments.

Keywords: sentiment analysis; feature importance; n-grams; drug reviews; mental disorders

Introduction

As product users increasingly utilize the internet to share and read content about other users' experiences, analyzing web-based datasets using sentiment analysis can help extract emotions and opinions from text. In particular, sentiment analysis can be used to improve patient treatment in the medical sector, being crucial especially with a rise in mental disorders as a consequence of the Covid19 pandemic (Bäuerle et al., 2020). Medical trials have been the main tool to ensure drug safety, but are difficult to access and comprehend for the regular patient (Adusumalli et al., 2015). Online reviews, on the contrary, are written in everyday language, are widely available, and reflect the opinions of a wider population (Fox & Duggan, 2013).

Textual online reviews are useful information sources, but the high dimensionality of the resulting datasets makes manual feature extraction a cumbersome task. Sentiment analysis, a linguistic approach that can predict polarity categories from text, is a valuable tool for analyzing large amounts of textual data. There are two main approaches of performing sentiment analysis: a linguistic rule-based approach and a statistical approach using Machine Learning. In this paper, the focus will be on the latter.

Sentiment analysis can identify the side effects of psychic drugs, which is crucial in improving the patient's life quality, as preventing severe side effects is vital for successful medical treatment (Sharif, Zaffar, Abbasi, & Zimbra, 2014; Stroup & Gray, 2018). By continuously keeping track of users' experiences with certain drugs, pharmaceutical companies could furthermore enhance drug safety in the post-marketing process (Alomar, Tawfiq, Hassan, & Palaian, 2020). Moreover, textual drug reviews can be used to better understand determinants of user drug satisfaction to recommend suitable drugs for specific diseases to users (Garg, 2021).

However, the explainability of Machine Learning models is one of the main problems in the field, as most algorithms act as "black boxes", making it impossible for humans to understand why an algorithm made certain predictions. Especially when the well-being of humans is involved, as it is the case for drug recommendation algorithms that decide which treatment someone receives, it is essential to get further insight into the model predictions (Burkart & Huber, 2021). In this study the most important textual features will therefore be analysed, using the Gini Index of Random Forest and the empirical log probability of Naïve Bayes, to analyse which features might improve the understanding of drug user satisfaction. Feature importance heavily relies on the selection of the Machine Learning model and parameters (Lai, Cai, & Tan, 2019), which is why a variety of different parameter settings will be explored.

Methods

Dataset

The used dataset contains user drug reviews obtained from the drugs.com website website at the end of June 2020, as provided at https://github.com/fzamberlan/scrapdrugs. The dataset consists of 24,833 entries of patients of five different diseases (ADHD, Anxiety, Bipolar disorder, Depression, and Schizophrenia). The effectiveness of these drugs is evaluated by drug users through written reports and a score ranging from one to ten. Zero-reviews were removed due to the inconsistency with the user's sentiment (as there were zero-score reviews with a positive textual evaluation). Based on the user scores, each review got assigned a positive (score between 1 and 4), neutral (score between 5 and 7), or negative sentiment score (score between 8 and 10), which were used as labels. The resulting dataset after these steps consisted of 11,207 entries, with 2,723 negative, 1,667 neutral, and 6,817 positive reviews (see Figure).

Preprocessing steps

During the preprocessing, the textual reviews were converted to lower-case, tokenized, and cleaned from punctuation, numbers and stopwords. We used the nltk English stopword list, a list of slang words, and a list containing the diseases and drugs to clean the data from words that add little meaning to the sentiment. Additionally, looking at the wordcloud, frequently occurring words, such as "side", "effect", and "medication" that occurred prevalently across all reviews were removed, resulting in stopword set 1. Stopword set

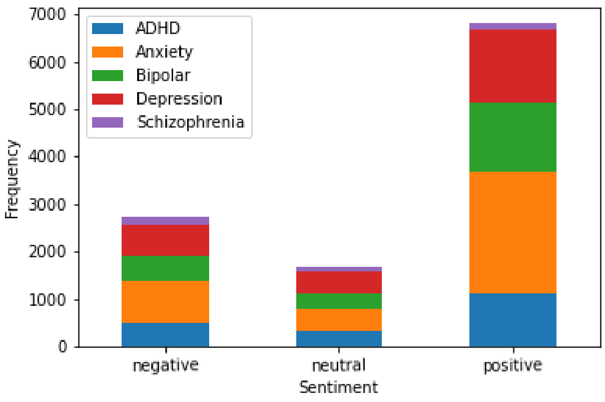

Figure 1: Frequency of reviews per sentiment and disease

2 additionally contained 'meaningless' words for unigrams (e.g.: "much", "gained" and "dont"), assuming that they only add meaning on higher n-gram levels (e.g.: "gained pounds", "much better"). A lemmatizer or stemmer was not used, as both yielded uninterpretable results. Three different n-gram settings: unigrams (one-word sequences), bigrams (two-word sequences), and a combination of unigrams and bigrams were furthermore tested to evaluate the effect of n-grams on the model performance.

The textual reviews, the features of interest, were then transformed into a numerical matrix, since Machine Learning algorithms require numerical input. For that purpose, we used the Term Frequency-Inverse Document Frequency (TF-IDF) which is a measure of importance of a word in a text. To calculate the TF-IDF, the Term Frequency (TF), which reflects the occurrence of a word in a text, and the Inverse Document Frequency (IDF), which measures the occurrence of a word in a document, were multiplied. When building the vocabulary, only terms which appeared in at least three documents were included (min_df = 3).

Implementation

Prior to training the model, the dataset was divided into the training and test set (80%-20%). As the dataset is highly imbalanced with a dominant positive class, an oversampling technique was used on the training set to reduce the bias towards the majority class. Two supervised Machine Learning classifiers, Random Forest and Multinomial Naïve Bayes, were trained for each disease, using six different feature settings (consisting of three n-gram and two stopword sets). Random Forest is a classifier that fits several independent decision trees and can prevent overfitting. For Random Forest, Grid Search with a 5-fold cross-validation was used to find the best performing hyperparameters for max_depth ([None, 2, 100, 500]), n_estimators ([100,500,1000]), and min_samples_split ([2,10,100]). Naïve Bayes, a probabilistic classifier, was trained without hyperparameters, as Naïve Bayes has a limited number of parameters.

Performance evaluation

For the model performance evaluation, we computed a confusion matrix, accuracy, precision, recall and f-scores. The weighted f-score, as a popular metric for imbalanced datasets, was used to compare the models. For each disease, the most important features, using the Gini index of Random Forest and the empirical log probability of Naïve Bayes, were extracted and visualized, to better understand the model predictions. The focus was on the latter as the empirical log probability allows to easily compare features across sentiment categories.

Results

This study investigated the effects of different textual features for the sentiment analysis of drug reviews using the drugs.com dataset. Textual reviews of patients of five mental disorders (ADHD, Anxiety, Bipolar disorder, Depression and Schizophrenia) were used to predict positive, neutral, and negative sentiments. Numerical scores, assigned by the users, were converted into three distinct sentiment categories (negative for 1-4, neutral for 5-7 and positive for 8-10). The dataset consisted of 11,207 entries after removing zero-scored reviews which did not accurately reflect the user's sentiment.

Before training the Machine Learning models, the reviews were cleaned and transformed into a numerical matrix, using the TF-IDF transformation. Since the data was imbalanced, an oversampling technique was applied on the training set. The performance of three different n-gram settings (unigrams, bigrams, unigrams+bigrams), two different stopword lists (with stopword set 2 additionally removing words such as "gained" or "dont", considered to be meaningless on an unigram level) and two algorithms (Naïve Bayes (NB) and Random Forest (RF) classifier) were compared. Table 1 shows the best model for each disease, based on the weighted F1-score. In the following analysis, the different parameter and feature settings will be compared, using the weighted F1-score. For all diseases, bigrams were outperformed by unigrams and unigram-bigram combinations. For ADHD, Anxiety and Depression, stopword set 1 achieved the best results while for the Bipolar and Schizophrenia group, stopword set 2 was preferred. Furthermore, the n-gram-range and stopword setting changed the gridsearch results, as visible in Figure 2 and Figure . The best model overall was the Bipolar model, using a Random Forest classifier with unigrams and stopword set 2, yielding the best performance for the positive class (F1-score=0.84) and the worst performance for the neutral class(F1-score=0.17), as shown in Table 3.

The five most important textual features for the model predictions were then extracted, visualized, and compared across n-gram and stopword settings. Figure 2 shows the most important unigrams for the Bipolar group, using stopword set 2. "Work" and "better" are the best features for the positive class, while "worse" and "horrible" are the best features for

Disease	Model	Precision	Recall	F.score	Accuracy	Stopword set	N-grams
ADHD	RF	0.65	0.71	0.67	0.71	1	unigrams
Anxiety	NB	0.72	0.65	0.68	0.65	1	unigrams+bigrams
Bipolar	RF	0.70	0.74	0.71	0.74	2	unigrams
Depression	NB	0.62	0.63	0.62	0.63	1	unigrams
Schizophrenia	NB	0.56	0.52	0.52	0.52	2	unigrams+bigrams

Table 1: Weighted performance metrics overview using best stopword setting results and n-grams

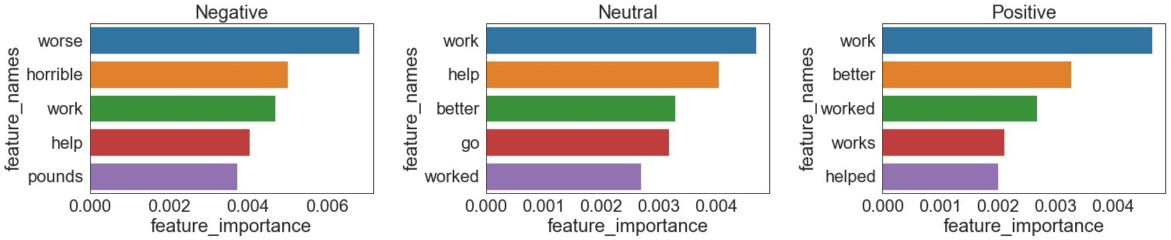

Figure 2: Most important features, using Naïve Bayes' empirical log probability, unigrams and the more restricted stopword set 2 for the Bipolar disorder group

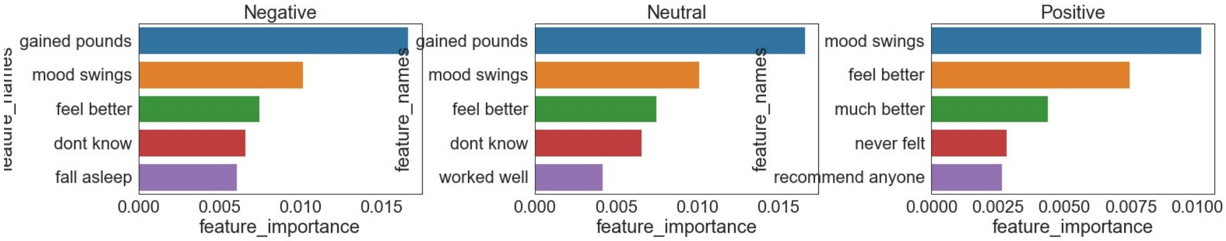

Figure 3: Most important features, using Naïve Bayes' empirical log probability, bigrams and the more restricted stopword set 1 for the Bipolar disorder group

stop	n-grams	max_depth	min_samples_split	n_estimators
1	(1,1)	None	10	500
2	(1,1)	500	2	1000
1	(2,2)	500	2	500
2	(2,2)	None	2	100
1	(1,2)	500	10	1000
2	(1,2)	None	2	1000

Table 2: Best parameter settings for Random Forest (stop = stopword set)

Class	Precision	Recall	F.score	Accuracy
negative	0.63	0.64	0.64	0.74
neutral	0.35	0.11	0.17	0.74
positive	0.79	0.89	0.84	0.74

Table 3: Performance metrics per class for the best model (Random Forest classifier for the Bipolar disease group, using unigrams and the more restricted stopword set 2)

the negative class. Figure 3 displays the most important bigram features, using stopword set 1. For predicting positive sentiment, ”mood swings” and ”feel better”, for predicting the negative class, ”gained pounds” and ”mood swings” were the best two features.

Discussion

For the Schizophrenia disease group, removing words assumed to be meaningless for unigrams (i.e.: ”gained”, ”much”) could improve the model performance for unigram-bigram combinations. This suggests that restricting the vocabulary seems sensible to improve the sentiment prediction when using single-word n-grams. For all diseases, unigrams and unigram-bigram combinations appear to be better textual features than bigrams as they enhanced the model performance. Furthermore, the results show that stopword sets should be separately defined for the five different diseases.

Using the drugs.com reviews, this study achieved a relatively low accuracy for predicting the sentiment. Although a resampling technique was applied on the training set, the model predictions for the positive majority class were better than for all other classes. The low F-scores show that the model could not accurately predict the neutral class. One main problem with the used data was that the numerical score,

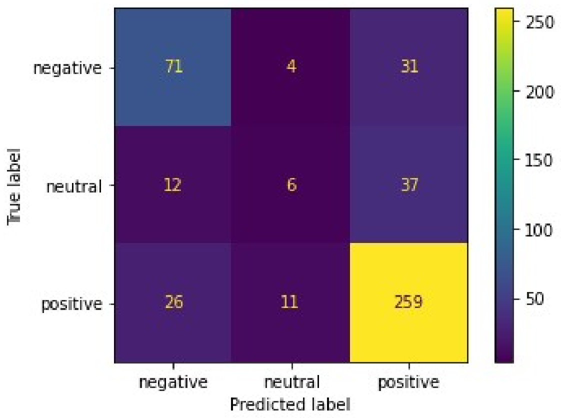

Figure 4: Confusion matrix for best model (Random Forest classifier for the Bipolar disease group, using unigrams and the more restricted stopword set 2)

used to train the classifiers, did not consistently align with the sentiment present in the textual review. In that way, some negative reviews included mostly positive words, possibly leading to large prediction errors.

In future research, binary classification could be tried, as the models could not accurately predict the neutral class. This study used disease-independent stopword sets only, disease-specific stopword lists could be therefore used to investigate the most important features. Moreover, inspecting the textual reviews manually in more detail and removing noisy reviews prior to the analysis might achieve a better model performance.

Conclusion

In this study, sentiment analysis was conducted comparing different textual input feature sets to better understand psychic drug satisfaction. This study demonstrates that careful restriction of the vocabulary and n-gram-ranges can heavily influence the prediction performances of Machine Learning models. The obtained low test results emphasize the importance to manually check the review quality beforehand to remove inconsistent reviews and better predict the sentiment. Overall, the results show that extracting disease-specific most important features enhances the model explainability and gives further insight into patients drug satisfaction. Thus, sentiment analysis with the integration of disease-specific knowledge could be used in future research to extract drug-related information from user reviews to improve the pharmaceutical treatment of mental disorder patients.

References

Adusumalli, S., Lee, H., Hoi, Q., Koo, S.-L., Tan, I. B., Ng, P. C., et al. (2015). Assessment of web-based consumer reviews as a resource for drug performance. *Journal of medical Internet research*, *17*(8), e4396.

Alomar, M., Tawfiq, A. M., Hassan, N., & Palaian, S. (2020). Post marketing surveillance of suspected adverse drug reactions through spontaneous reporting: Current status, challenges and the future. *Therapeutic advances in drug safety*, *11*, 2042098620938595.

Bäuerle, A., Teufel, M., Musche, V., Weismüller, B., Kohler, H., Hetkamp, M., ... Skoda, E.-M. (2020). Increased generalized anxiety, depression and distress during the covid-19 pandemic: a cross-sectional study in germany. *Journal of Public Health*, *42*(4), 672–678.

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, *70*, 245–317.

Fox, S., & Duggan, M. (2013). Health online 2013. pew research center. *National survey by the Pew Research Center's Internet and American Life Project*.

Garg, S. (2021). Drug recommendation system based on sentiment analysis of drug reviews using machine learning. In *2021 11th international conference on cloud computing, data science & engineering (confluence)* (pp. 175–181).

Lai, V., Cai, J. Z., & Tan, C. (2019). Many faces of feature importance: Comparing built-in and post-hoc feature importance in text classification. *arXiv preprint arXiv:1910.08534*.

Sharif, H., Zaffar, F., Abbasi, A., & Zimbra, D. (2014). Detecting adverse drug reactions using a sentiment classification framework.

Stroup, T. S., & Gray, N. (2018). Management of common adverse effects of antipsychotic medications. *World Psychiatry*, *17*(3), 341–356.