Sentiment Analysis: Gender Bias in Machine Learning Classifiers

Yuqi Deng

1. Introduction

Today, people usually use the internet to get information about healthcare and also to comment on the care they receive [1]. In this study we are mainly dealing with a sentiment analysis about the patient's satisfaction or dissatisfaction with the doctor from their unstructured comments. We did this by analyzing the existing comments of the patients so that we could know whether they had a positive or negative sentiment towards their doctors. In addition, the primary research question that guided our research was whether there is a gender bias in the classifiers generated through machine learning.

Our dataset comes from ratemds.com, a platform where reviews of clinicians can be made. Patients can freely express their opinions about their doctors at the platform. Therefore, this platform has huge and rich data about doctors' reviews. The main data for this research comes from two papers, Wallace et al. (2014) and López et al. (2012).

In this project, we will analyze the dataset by using One-R (One Rule) model as the baseline algorithm and Decision Tree and Bagging models as the additional classifiers. Our aim was not only to assess unstructured comments, but also to explore the potential gender bias present in models for a more unbiased sentiment analysis.

2. Literature review

In a report entitled "What Patients Say About Their Doctors Online: A Qualitative Content Analysis", the authors analyzed 712 comments on physician rating websites. They focused their analysis on 445 primary care physicians from different urban areas in the United States. The main aim was to explore the content and themes of the online reviews. The results of the study showed that 63% of the reviews were positive. And there were differences between overall comments and specific descriptions. In the overall comments patients simply praised the doctor. In the specific descriptions, 69% of the patients gave positive comments about the doctor's interpersonal abilities. 80% of the patients gave positive comments about the doctor's technical abilities. Patients gave the doctor mixed ratings on systems issues, including appointment scheduling and telephone visits. The study also emphasized that other factors such as staff behavior, access and convenience were also present in patients' ratings of their doctors. Overall, this study provides insights into online reviews so that patients can have a better understanding of the doctors being rated.[2]

The report, entitled "Harnessing the Cloud of Patient Experience: Using Probabilistic Models of Text to Analyze Online Physician Reviews" explored the emotions that are latent in physicians' online reviews about different aspects of care . An empirical evaluation of approximately 60,000 reviews was conducted using probabilistic models of text. In addition, the authors propose a text generation model based on f-LDA. It was able to capture the potential sentiment of the reviews on each care aspect. By adding the f-LDA output to the regression model, its correlation with state-level measures of healthcare quality was improved, especially in predicting outcomes such as patients' visits to primary care physicians after hospital discharge. This study shows that the potential for online physician review analysis can be increased by combining large-scale quantitative modelling with traditional qualitative analysis.[1]

In a report entitled "Impact of Surgeon Gender on Online Physician Reviews", the authors explore and analyze whether there are differences in patients' online reviews of male and female surgeons, particularly in terms of content and quality. Using an deliberate sample of physician rating platforms RateMDs.com and Yelp.com, this report focus on 431 reviews of the top 20 surgeons from four of the most densely populated metropolitan areas in the United States (New York, Houston, Los Angeles, and Chicago), including overall ratings, communication,

technical ability, and supportive elements. The results showed that the most comments were positive, and there was no difference in overall ratings by gender. However, females rated higher in the social interaction part, while males rated higher in the technical skills part.[3]

In this report, entitled "Unhappy Patients Are Not Alike: Content Analysis of the Negative Comments from China's Good Doctor Website", the authors explored the negative patient comments on China's Good Doctor website with the main aim of helping doctors to improve patient satisfaction. By analyzing 3012 negative comments from 1029 doctors from five famous hospitals in Beijing, it was shown that patient dissatisfaction exists in different aspects of the consultation process. The highest number of negative reviews were received from doctors at Obstetrics and Gynecology (20.12%) and Internal Medicine (16.17%). The highest rankings came from Dermatology and Sexually Transmitted Diseases (mean 5.72) and Andrology (mean 5). The negative comments were usually focused on low consultation time (19.16 %), doctor impatience (17.50%) and poor treatment (12.28%). Special groups such as patients accompanying elderly patients or children have a low tolerance for medical care. This analysis finds the elements of negative comments for doctors, and the results of the research can help them to improve their healthcare services.[4]

3. Method

In this section, it will apply the pre-processing and feature selections on the data. Explore models with better performance and apply them to unbalanced dataset.

3.1. Pre-Processing

I removed the 'Unnamed: 0' column from the 'train_tfidf', ‘train_embedding’, ‘validation_tfidf’, and ‘validation_embedding’ DataFrames by using the ‘drop’ method with the ‘inplace=True’ parameter to make the DataFrames more clean and more focused on relative feature as he column is an index column without any meaningful information here.

3.2. Feature Selection

The feature selection on word embeddings and TFIDF using Chi-squared and Mutual Information methods with various feature counts (k). It normalizes the features and evaluates a Decision Tree model's accuracy for each combination of k and feature selection method. This process aids in identifying the optimal feature subset for the model and potentially enhancing predictive performance. Although the dataset is large the difference in their performance is small.

3.3. Model Selection

In this section, I will focus on the performance of different models and the reason of the better performance.

3.3.1 Baseline Model

One R model is a simple and easy to understand with straightforward algorithm, it is one of the good choices for baseline model. Since the cost of training a One R model is small, this is well suited for our large datasets. Thus, it can be a basis for more complex modelling. However, it also has limitations. This is because it may oversimplify complex relationships in the data and be sensitive to uncorrelated features. If advanced models do not perform as well as One R model, the reason may be that the problem is simple, or the data is not complex enough to support more advanced model use.

3.3.2 Decision Tree Model

Decision trees suitable for a wide range of features in medical reviews as it can handle both categorical and numerical data. Its hierarchical structure can help to discover complex relationships and interactions in the data. However, decision trees are easily overfitted, and can affect the results especially if the dataset contains noise. Comparison with a baseline (One R Model) can show that whether the decision tree improves the prediction performance and explores the complexity of the dataset.

3.3.3 Bagging Model

Bagging model based on Decision Tree, which enhances predictive performance by training multiple decision trees on bootstrap samples, which reduces overfitting. This results in more stable predictions due to reduced variability. It is also suitable for large data. Its combined methods can help to compensate for the overfitting disadvantage

associated with decision trees. However, its interpretability may be reduced compared to a single decision tree. The bagging model may result in higher computational costs and longer computational time due to training multiple models.

3.4. Evaluation

In this evaluation, three classifiers (One R, Decision Tree (DT), and Bagging) were examined using TFIDF and word embedded datasets for sentiment analysis of patient comments. Firstly, the accuracies of the three classifiers were calculated separately. However, the distribution of the training and validation data was unbalanced. However, because of the unbalanced distribution of the training and validation data, in which the number of ratings of 1 is significantly greater than the number of ratings of -1, accuracy is not enough to make an evaluation well. It is necessary to use some more metrics that can give more weight to the minority class.
To address this, the performances of the classifiers on both TFIDF and word embeddings datasets were further evaluated using the evaluation metrics such as precision, recall and F1 score calculated by the 'precision_recall_fscore_support' function as well as the macro-averages. The focus of this part is to compare the performances of these models in capturing the sentiment expressed from patient reviews. Precision assesses the accuracy of positive predictions, Recall measures the ability to recognize positive sentiment, and the F1-score provides a balanced measure that considers both metrics. Models are compared based on these values to see how well they handle the complexity of analyzing sentiment about healthcare reviews. This method gives a better evaluation, especially for unbalanced datasets.
Moreover, the confusion matrices are used to provide a detailed and visual understanding of the performance of the classifiers especially for unbalanced data. They provide a visualisation of these performances in terms of classifying sentiment categories.

4. Result

4.1. Unbalanced datas

The number of data with a rating of 1 in the training data is about 2.55 times the number of data with a rating of -1. And in the validation data, there is about 2.76 times (see Figure.1). Therefore, accuracy is not enough for evaluation.

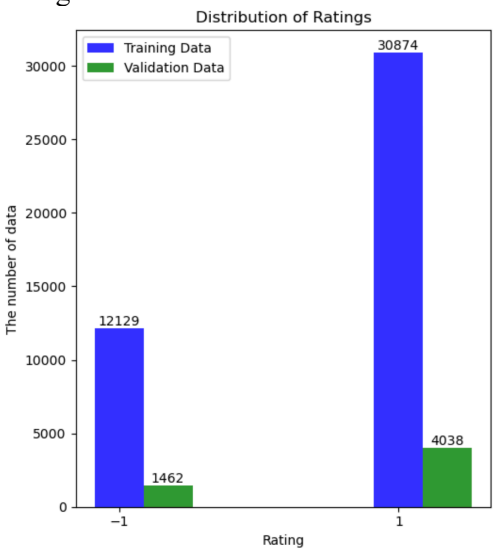

Figure 1. The bar chart of the datasets

4.2. Evaluation of all models

The Table.1 and Table.2 show that Bagging model has the best performance in all metrics and datasets.

Model	Accuracy	Precision	Recall	F1-score
One R	0.7460	0.6904	0.7188	0.6991
Decision tree	0.8118	0.7590	0.7577	0.7584
Bagging	0.8645	0.8278	0.8225	0.8251

Table 1. The Evaluation of word embeddings dataset

Model	Accuracy	Precision	Recall	F1-score
One R	0.7653	0.7797	0.5714	0.5616
Decision tree	0.8378	0.7907	0.8082	0.7985
Bagging	0.8724	0.8327	0.8511	0.8410

Table 2. The Evaluation of TFIDF dataset

4.3. Confusion Matrixes

As can be seen from Figures 4.1-4.3, most of the predictions are correct. The ratio between the values of (1,

1) and (-1, -1) is similar to the ratio between rating of 1 and -1 in the original data. The performance of bagging model is the best.

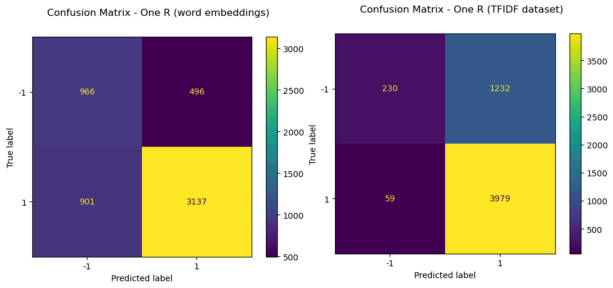

Figure 4.1. Confusion matrix of one R model

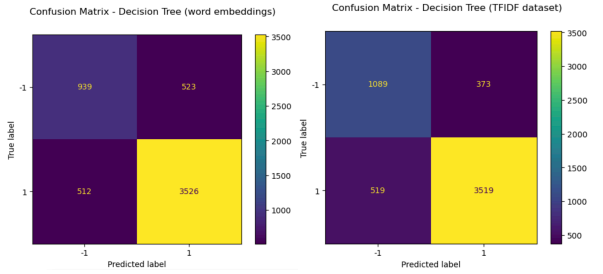

Figure 4.2. Confusion matrix of decision tree model

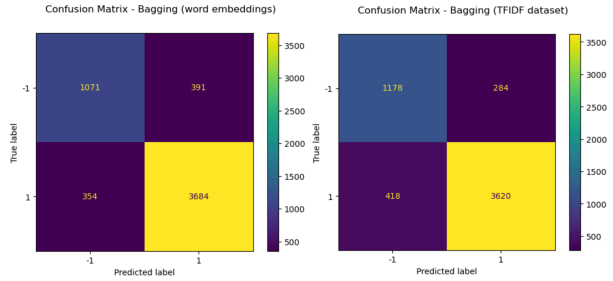

Figure 4.3. Confusion matrix of bagging model

4.4. Feature Selection

For the word embeddings datasets (See Figure 2), it has the best performance for both methods at k = 100. For TFIDF (See Figure 3), its accuracy increases as k increases and the best one is “200” .

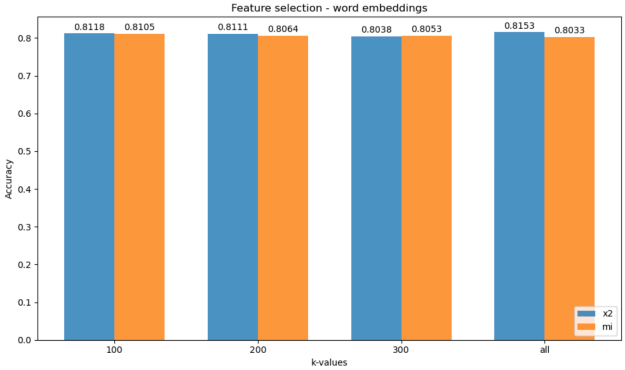

Figure 2. Word embeddings dataset feature selection

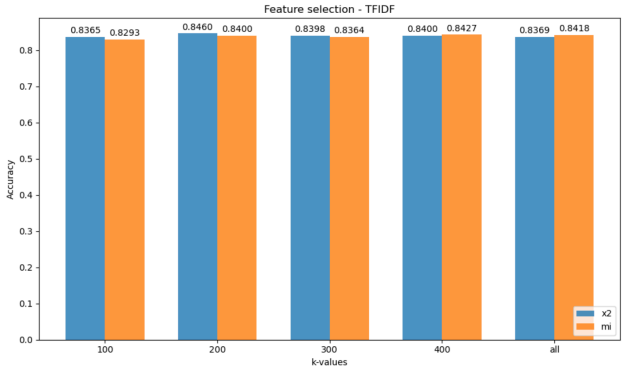

Figure 3. TFIDF dataset feature selection

4.5. Learning Curve

From Figures 5, the chosen depth is 11 for all the datasets with word embeddings. The chosen depth for all the datasets with TFIDF is 46. Because both datasets are good with higher accuracy at this point (See Figure 5.1-5.2, Table 3).

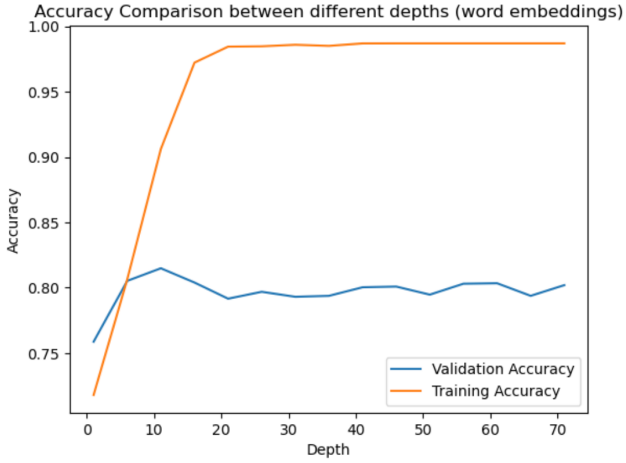

Figure 5.1. Decision tree learning curve (word embeddings)

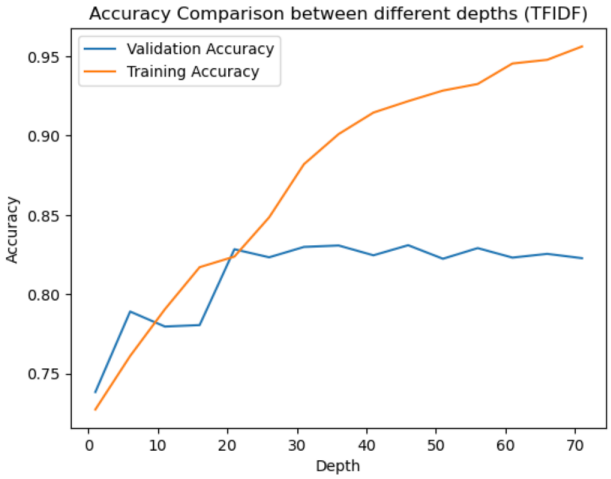

Figure 5.2 Decision tree learning curve (TFIDF)

Dataset	Model	Accuracy
Word embeddings	Decision tree	0.8291
	Bagging	0.8680
TFIDF	Decision tree	0.8360
	Bagging	0.8724

Table 3. The Accuracy after adjustment parameters

5. Discussion

Without considering the gender, bagging model with TFIDF dataset performs the best of all (See Table 1-3). The Bagging model is a classifier with almost no gender bias as performs very fairly in the dataset of the word embedding representation with different genders. It is also accurate for both males and females in the dataset of the TFIDF representation, but it performs poorly in the dataset for gender unknown. The One R model seems to be the most unfair compared to the other two models. (See Figure 6.1-6.2). The TFIDF vocabulary contains some unnecessary information that affects accuracy such as the number '100'. In addition, since the number of unknowns is much smaller than the number of male and female, this is more likely to lead to bias in accuracy. Overall, male's will be slightly more accurate than female's, probably because male has about 2.39 times as many comments as females.

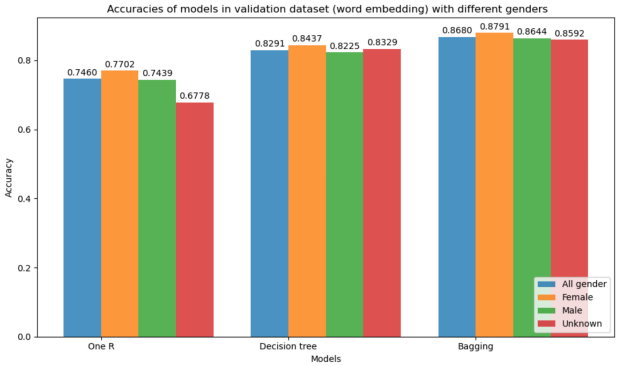

Figure 6.1. Accuracies (TFIDF) of 3 models with different genders

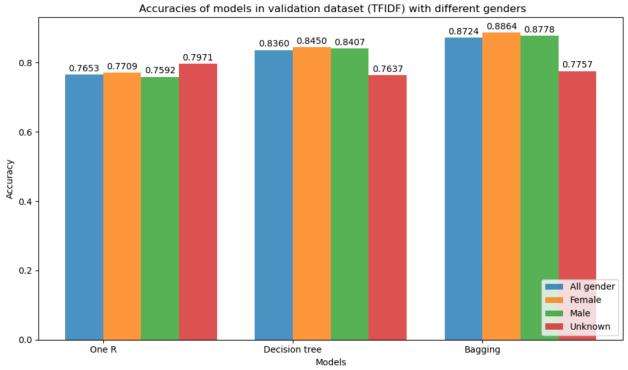

Figure 6.2. Accuracies (TFIDF) of 3 models with different genders

6. Conclusions

In summary, this study analysed the sentiment of a dataset about patient reviews for doctors on ratemds.com by using three different machine-learning classifiers. The Bagging model, especially when using TFIDF data, shows excellent sentiment analysis ability. However, subtle differences were found in the "unknown" gender. The One R model is a relatively unbiased model because it is too simple to deal with the complexity of the relationships in the data. The unbalance in the data set brings a challenge to the research. Future research could attempt to explore more natural language processing methods on a more refined dataset. This research is convenient for healthcare people and highlights that more methods can be explored to reduce potential bias in sentiment analysis.

References

[1] Wallace, B. C., Paul, M. J., Sarkar, U., Trikalinos, T. A., and Dredze, M. (2014). A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. J Am Med Inform Assoc, 21(6):1098–103.

[2] L´opez, A., Detz, A., Ratanawongsa, N., and Sarkar, U. (2012). What patients say about their doctors online: a qualitative content analysis. J Gen Intern Med, 27(6):685–92.

[3] Marrero, K., King, E., & Fingeret, A. L. (2020). Impact of Surgeon Gender on Online Physician Reviews. The Journal of surgical research, 245, 510–515. https://doi.org/10.1016/j.jss.2019.07.047

[4] Zhang, W., Deng, Z., Hong, Z., Evans, R., Ma, J., & Zhang, H. (2018). Unhappy Patients Are Not

Alike: Content Analysis of the Negative Comments from China's Good Doctor Website. Journal of medical Internet research, 20(1), e35. https://doi.org/10.2196/jmir.8223