Supervised Learning Techniques for Gender Bias problem on Clinician Reviews

Anonymous

1. Introduction

Ratemds.com is a clinician review website in United State. It's a wonderful website for people to express their individual opinions and thoughts about their doctor. Individual discussions will be posted online with feedback and help. However, not all reviews are positive and helpful, some patient may post negative opinions to doctor due to unsuccessful treatment, over charging prices or other reasons. Both positive and negative comments are reflect to professional ability of specific doctor, and valuable data to demonstrate rating of doctors. However, these reviews might influenced by several factors, gender, is one of them.

In this project, different supervised machine learning techniques were developed to classify positive and negative comments contained in all review data sets. The data sets used in this project is derived from qualitative work conducted by L´opez et al [3] and further developed into a data set suitable for computational analysis by Wallace et al [5]. Data set contains both labeled and unlabeled reviews from both male and female. The data sets used in training and development are labeled with rating and gender. The total number of instances of training/validation/test/unlabelled in all data set is listed in Table 1 and Table 2.

This project aims to address the following research question: **How does gender bias and ratio influence the performance of different supervised machine learning techniques.**

To answer this research question, lists of data for different gender is separated from original data set, and machine learning techniques leveraging different gender lists are implemented. In this project, four supervised methods were analysed (not including baseline), respectively Naive Bayes (NB), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP). The performance o both combined gender and splitted gender data sets will be tested by these methods to investigate whether gender bias influence the performance.

Training	Validation	Test-Unlabelled
43,003	5,500	5,514

Table 1. Count of comments in Original data set.

	Training	Validation	Test
Embedding	272	272	272
TF-IDF	582	5,500	5,514

Table 2. Instances in Embedding/TF-IDF data set.

2. Literature Review

Over the recent years, a burgeoning community of researchers have turned their focus towards identifying and mitigating gender bias prevalent in social platforms. This is in line with the broader societal recognition of gender equity issues, mirroring a collective aspiration to cultivate a more inclusive digital environment. In a notable development in 2023, a paper authored by by Kaur et al [2] ideates a method of reduction gender bias in social with machine learning technique that based on multi-layer perceptron due to their high performance. Christine Basta et al [1] pointed out that word embeddings have been proved to emphasis the bias present in data source. Furthermore, Sandro et al [4] adjusted LR algorithm to mitigate unwanted discrimination such as gender.

3. Method

3.1. Data Introduction and Pre-processing

The raw data set published on Kaggle has a column named 'review-text-cleaned' that contains positive and negative comments to doctors, which can not be used as input to classification algorithm. To resolve this

problem, I implemented a function to convert text to ASCII value. The entire data set contains 3 different part: original data, embedding and TF-IDF. Original data has three sub-set: Train, Validation and Test with no Labels. Both embedding and tf-idf set contain train, validation and test. For any sub-sets that contain column 'review-text-cleaned', a new column 'ascii-sum' will be added, the value is equivalent to sum of ASCII value in 'review-text-cleaned' of row. By applying this method, string should be converted into int/float value, and higher ASCII value means longer string being converted, also means stronger tone expressed by patient. Files are read without first column.

3.2. Baselines

The baseline methods employed in this study include the Zero Rules (Zero-R) and Random baseline. The Zero-R classifier predicts an instance based on the most frequent class in the training data. On the other hand, the Random classifier operates by indiscriminately predicting labels randomly, providing a average measure of performance. In this project, the average accuracy score will be calculated after performed ten round of Random baseline.

3.3. Supervised Learning Methods

Naive Bayes(NB): This classification method operates on the foundation of Bayes' theorem, with the simplifying assumption that each feature pair behaves independently. There are two NB methods used in this project: Gaussian Naive Bayes(GNB) and Multinomial Naive Bayes(MNB). GNB assumes that continuous features adhere to a Gaussian distribution, making it ideal for data sets with continuous attributes.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

On the other hand, MNB is tailored for discrete data, like word counts in text, under the premise that such features follow a multi-nomial distribution.

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

Logistic Regression(LR): LR is a rather simple and efficient learning technique. It predicts the probability that a given instance belongs to a particular category by fitting data to a logistic function. Despite the fast training speed and high accuracy, LR is very likely influenced by over-fitting, especially when training with discrete data. In this project, different 'weight' and 'max-iter' are tried to find best performance.

K-Nearest Neighbors(KNN): KNN is a non-parametric and distance-based learning method. KNN classifies a data point based on distance and K number of neighbors around data point. This gives KNN incredibly fast training time, but relatively slow of predicting time. In this project, different number of neighbors are tried to find best performance.

Multi-Layer Perceptron: MLP is a type of artificial neural network. It is used to model complex, non-linear relationship in data source. Each neuron processes input and applies an activation function, then pass the result to next layer. It is very suitable for this data set to model non-linear relationship. In this project, different 'hidden-layer-sizes' and 'learning-rate' will be used in different data subset to test best performance. However, due to the reason of extremely long training time in MLP classifier, I am unable to fully test all possible combinations of hyperparameters, instead, I will demonstrate the best result in data set.

3.4. Evalution Metrics

Four metrics will be applied in performance result. The performance before separate gender and after separated gender will illustrate individually. All four metrics will demonstrate average value after combine positive(1) and negative(-1) ratings.

- Precision(P): It is the ratio of correctly predicted positive observations to the total predicted positives

- Recall(R): Measures the proportion of actual positives that were correctly identified by a classification model

- Accuracy(Acc): The ratio of correctly predicted observations to the total number of observations

- F1-Score(F1): A metric that combines precision and recall into a single value, providing a balanced measure of a model's performance

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

	TFIDF				Embedding				Origin			
	Acc.	P	R	F1	Acc.	P	R	F1	Acc.	P	R	F1
Zero-R	0.71	0.365	0.5	0.425	0.73	0.365	0.5	0.425	0.73	0.365	0.5	0.425
Random	0.5	0.505	0.505	0.47	0.51	0.515	0.515	0.68	0.50	0.505	0.503	0.47
GNB	0.86	0.815	0.865	0.835	0.87	0.825	0.875	0.84	-	-	-	-
MNB(alpha 0.1)	0.88	0.88	0.815	0.835	-	-	-	-	-	-	-	-
LR(800 iter)	0.91	0.875	0.89	0.995	0.92	0.9	0.89	0.9	0.27	0.135	0.5	0.21
LR(weight 1:0.4 01:0.6)	0.90	0.87	0.895	0.88	0.91	0.885	0.91	0.895	0.27	0.125	0.5	0.21
KNN(9 neighbors)	0.84	0.81	0.765	0.785	0.88	0.87	0.82	0.84	0.73	0.365	0.5	0.425
MLP(300 hidden layer)	-	-	-	-	0.92	0.89	0.895	0.89	-	-	-	-
MLP(adaptive rate)	-	-	-	-	0.92	0.895	0.89	0.89	-	-	-	-

Table 3. Performance before splitting gender.

	TFIDF				Embedding				Origin			
	Acc.	P	R	F1	Acc.	P	R	F1	Acc.	P	R	F1
Male GNB	0.88	0.89	0.81	0.835	0.87	0.835	0.88	0.85	-	-	-	-
Female GNB	0.87	0.83	0.88	0.845	0.91	0.815	0.88	0.85	-	-	-	-
Male MNB(alpha 0.1)	0.88	0.89	0.81	0.835	-	-	-	-	-	-	-	-
Female MNB(alpha 0.1)	0.89	0.87	0.84	0.855	-	-	-	-	-	-	-	-
Male LR(800 iter)	0.91	0.895	0.89	0.89	0.92	0.9	0.905	0.905	0.27	0.135	0.5	0.215
Female LR(800 iter)	0.92	0.895	0.895	0.895	0.92	0.895	0.915	0.895	0.26	0.13	0.5	0.21
Male KNN(9 neighbors)	0.85	0.83	0.78	0.8	0.88	0.87	0.82	0.84	0.73	0.365	0.5	0.42
Female KNN(9 neighbors)	0.84	0.81	0.765	0.785	0.89	0.865	0.845	0.855	0.74	0.37	0.5	0.425
Male MLP(300 hidden layer)	-	-	-	-	0.91	0.9	0.88	0.89	-	-	-	-
Female MLP(300 hidden layer)	-	-	-	-	0.93	0.905	0.915	0.91	-	-	-	-

Table 4. Performance after splitting gender.

The origin and embedding data set is not used in GNB and MNB because they contain negative values, this will rise an error when training model. The origin data set also not used when training GNB, this is because despite converted string to int/float value, it will cause over-fitting due to all the values are unique. The over-fitting problem remains even after normalization (min-max normalization) applied to data set.

$$X_{\text{norm}} = a + \left(\frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \right) \cdot (b - a)$$

The over-fitting problem also appeared when training origin data set in LR model, but it can be resolved by applying regularization and adjust penalty hyperparameter. TFIDF and Origin sets are not used in MLP model. This is because the training process of MLP is extremely time consuming, it will take whole day to observer the different combination of hyperparameters with data sets. Therefore, I only used embedding data set to compare the performance before and after gender splitting.

4. Results and Discussion

In terms to answer research question addressed in section 1, I implemented four different supervised machine learning methods: NB, LR, KNN and MLP. Then observed the performance of these four techniques under different gender condition. All algorithms are implemented by using sk-learn library. Each techniques are explored as many combination of data sets and hyperparameters as possible, intended to maximize the output results. The performance results of both before and after splitting gender are listed in Table 3 and Table 4.

	male	female	male-female ratio
Training	27511	12141	2.26 : 1
Validation	3584	1497	2.39 : 1

Table 5. number of Male and Female instances

4.1. Influence on metrics when split gender

As performance presented in Table 3, all results under TFIDF and Embedding sets are better than baseline classifier. LR (800 iter) model consistently outperforms other models across both TFIDF and Embedding data sets. Only origin set are worse than baseline in LR and KNN, with LR notably dropping in performance in both accuracy and F1 score. I suspect this is caused by the reason that origin set are not meant to use as input directly. Even after normalization process applied, there is still not much features for model to learn.

I compared the metrics results in Table 4 with Table 3, I found out that there are little differences in performance between male and female data sets. Female data set are commonly slightly better than male data set (less or equal than 0.1). MLP surprisingly worse than LR in overall result. However, LR might be influenced by over-fitting problem, whereas MLP shouldn't be worried about this problem.

4.2. Influence on methods when split gender

After gender splitted, both male and female in all models have slightly increase on result performance. I intend to know that whether the results after gender split are affected by gender ratio in origin data set. To answer this question, I calculated gender ratio in origin data set shown in Table 5. The Table 5 demonstrates that the number of male instances are more than twice of number of female instances. This suggests that, the female results performed by all models after gender split, may not as 'complete' as results produced with male data sets. Hence, even though results produced by female data sets are slightly higher than male results, but it's not as reliable as results produced by male data sets.

I also investigate the positive-negative comments ratio. The purpose is to find out whether positive-negative comments ratio is influenced by gender. This will reflect to performance after gender split, because P, R and F1 in metrics is the average value of positive(1) and negative(-1) comments. The result illustrated in Figure 1. Figure 1 points out that the positive-negative comments ratio is similar in both male and female. This suggests that comments published by different gender in this data set does not affect the data balance. Hence, the results produced by both gender are safe and reliable from 'positive-negative ratio' imbalance, and models used in this project are not influenced by imbalanced ratio of comments.

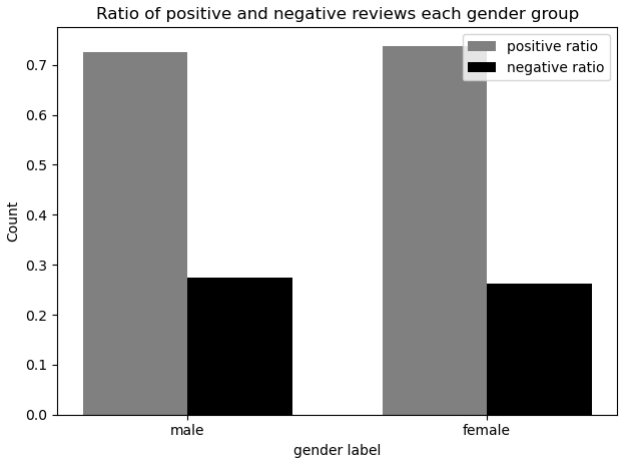

Figure 1. Comments Ratio in Gender

5. Conclusions and Future Improvement

In this project, I implemented four different supervised machine learning techniques, namely NB, LR, KNN and MLP. I explored as many combinations of data sets and hyperparameters in different as possible, to guarantee the best performance of model. Furthermore, to answer the research question, I splitted data sets into male and female subsets, and retrain these subsets to compare with original results. It is observed that performance from before and after gender split doesn't have significant impact on performance in different model. The conclusion can be dropped that performance produced by these techniques is not influenced by gender bias factors in these data sets. However, there are still two thing worth noting: First, number of male instance is much greater than female instance. This could cause model to over-learn male patterns and ignore female patterns. Second, even though

the positive-negative comments ratio are equivalent in male and female data sets after gender splitting, it is not balanced inside of male and female data sets. Figure 1 demonstrates that number of positive reviews are greater than negative reviews. This means model will get high F1 score when predicting positive reviews, and produce relatively low performance when predicting negative reviews. To improve these problems and make models more robust, simply adjust ratio of both male-female data sets and positive-negative comments.

6. Ethics Statement

The data used in this project were publicly available when collected on Kaggle. To protect the privacy and anonymity of the users in the data set, all reviews provided in data set were without name. All processed data and results will not released online without authorized and will not used in any places except this project.

References

[1] Christine Basta, Marta R Costa-Jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.08783*, 2019.

[2] Jaskirat Kaur, Sanket Mathur, Sukhpreet Kaur, Anand Nayyar, Simar Preet Singh, Sandeep Mathur, et al. Evaluating and mitigating gender bias in machine learning based resume filtering. *Multimedia Tools and Applications*, pages 1–21, 2023.

[3] Andrea López, Alissa Detz, Neda Ratanawongsa, and Urmimala Sarkar. What patients say about their doctors online: a qualitative content analysis. *Journal of general internal medicine*, 27:685–692, 2012.

[4] Sandro Radovanović and Marko Ivić. Enabling equal opportunity in logistic regression algorithm. *Management: Journal of Sustainable Business and Management Solutions in Emerging Economies*, 28(2):55–66, 2023.

[5] Byron C Wallace, Michael J Paul, Urmimala Sarkar, Thomas A Trikalinos, and Mark Dredze. A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *Journal of the American Medical Informatics Association*, 21(6):1098–1103, 2014.