# **Introduction**

**Investigating Gender Bias in Sentiment Analysis Classifiers**

Anonymous

This project seeks to analyse the efficacy of various machine learning models in discerning patient sentiments towards clinicians using patient reviews. A key area of exploration is whether these models exhibit any gender bias when predicting sentiments.

The dataset encompasses 54,107 reviews from patients about 19,097 clinicians, where sentiments are binary: -1 indicates a negative sentiment, and 1 indicates a positive sentiment. Within this dataset, 72% of the reviews reflect a positive sentiment. For training, a subset of 43,003 reviews is used, complemented by a validation set of 5,500 reviews. An additional 5,514 reviews without sentiment labels constitute the test set.

To facilitate the use of textual reviews in machine learning models, the data has undergone pre-processing. Techniques employed include term frequency-inverse document frequency (TF-IDF) (Manning, 2009) and Sentence Transformer Embeddings (Reimers and Gurevych, 2019), which transform the text into numerical vectors.

The primary objective is to train models using the training dataset and gauge their performance on the validation and test sets. Models demonstrating superior performance on the entire dataset will subsequently be assessed for any potential bias when applied specifically to male and female demographic groups within the data.

# **Literature Review**

## **Dataset**

The datasets utilized for this project are sourced from the studies of Wallace et al. (2014) and López et al. (2012), which concentrated on patient reviews from Rate MDS website, a notable clinician review platform in the United States. The principal aim of their research was to delve into online clinician reviews to identify prevailing patient sentiments towards healthcare. It's important to note from their work that there was no differentiation made among types of medical care, despite the likely variance in drivers of positive sentiment among them.

## **Models**

Prior research by Pang, Lee, and Vaithyanathan (2002) has demonstrated that Sentiment classifier performance has been to be topic-dependent, indicating that no single classifier universally excels. Among the various classifiers, Support Vector Machines (SVM) and Multilayer Perceptron (MLP) have consistently stood out across different domains.

Guia et al. (2019) explored four machine learning algorithms—Naive Bayes, SVM, Decision Trees, and Random Forest—for sentiment analysis on the Amazon Reviews: Unlocked Mobile Phones dataset. Of these, SVM stood out, achieving around 89% on key metrics such as Accuracy, Precision, Recall, and F1 score. While Random Forest had notable results, Decision Trees and Naive Bayes variants were less impressive. Given SVM's robust performance in this research, it is a prime choice for comprehensive sentiment analysis tasks.

In another study on online shopping reviews, Singla et al. (2022) compared MLP and Multinomial Naïve Bayes using data from Flipkart and Amazon. MLP had better accuracy (98.89%) than MNB (93.96%). They attributed the superior performance of MLP to its ability to model complex, non-linear relationships in the data, making it more adaptable to varied patterns in reviews. Based on their findings, MLP appears especially effective for extensive review sets, highlighting its effectivity for sentiment analysis tasks.

## **Gender Bias**

Research conducted by Bolukbasi et al. (2016) highlighted that machine learning (ML) and artificial intelligence (AI) models often inherit societal bias when trained on human-produced texts. This corroborates Datta et al.'s (2015) observations that models, when trained on past data exhibit gender bias.

In a detailed analysis, Kiritchenko and Mohammad (2018) uncovered that over 75% of sentiment analysis systems exhibited biases, favoring one gender or race over another. They raised concerns about the potential implications of even a consistent small bias in downstream applications. Their findings emphasized that the bias can be different depending on the specific affect dimension involved and that biases related to race appeared more prevalent than those related to gender. This indicates a pressing need to address these biases in the design and training of ML systems.

# **Method**

## **Dataset**

The pre-processed datasets with two feature representations, TF-IDF and embeddings, were employed on various models to determine which performed optimally based on the evaluation metrics. The superior representation was then used for gender-specific evaluation to delve into the potential presence of bias.

To carry out the gender-specific evaluation, the row indexes of the raw dataset corresponding to each gender were utilised to extract the instances from the TF-IDF and embeddings datasets for the respective gender. For this project, the gender-specific evaluation was concentrated on the 'Male' and 'Female' categories, excluding the 'Unknown' gender category.

## **Models**

### **Baseline**

ZeroR, the chosen baseline, predicts the dataset's most common label. Although ZeroR has been evaluated for the raw dataset, the performance for both TF-IDF and embeddings representation would be the exact same. Despite its simplicity, it benchmarks other models and checks if they capture data patterns or just replicate the class distribution.

### **Classifiers**

SVM is known for its effectiveness in sentiment tasks. It works especially well with text data because it can handle lots of information (Guia et al., 2019). This makes it ideal for separating different sentiments in the dataset.

MLP, a kind of neural network, is recognized for its ability to understand complex relationships (Singla et al., 2022). With its many connected layers, MLP can pick up and learn detailed patterns in data, making it useful for understanding sentiments expressed as different feature representations.

### **Evaluation metrics**

The primary metric used to gauge the effectiveness of the models was accuracy. This metric measures the proportion of correctly predicted sentiment outcomes in the overall dataset.

Additional metrics extracted from the Classification Report were also utilised to provide a more nuanced assessment of the model's performance. These include:

1. Precision: Indicating the proportion of true positives among all positive predictions, which is significant when the cost of false positives is high.

2. Recall: Measuring the proportion of true positives among all actual positives, which is crucial when missing a true positive is costly.

3. F1-Score: The weighted average of Precision and Recall, providing a balanced measure especially useful when both false positives and false negatives carry significance.

The use of these metrics alongside accuracy should provide a more thorough understanding of a model’s performance, accounting for not only correct predictions but also understanding the model's capability in managing false identifications.

# **Results**

## **Entire Dataset**

### **Initial Observations**

A bar graph with blue squares

Description automatically generated

Figure 1. The distribution of ratings for the training dataset

The training dataset demonstrates a skew towards positive ratings, as evidenced by the count of positive ratings (1) standing at 30,874 and negative ratings (-1) at 12,129. This skew may impact the performance of the models, as well as indicates the presence of class imbalance in the dataset.

### **Accuracy**

|  |  |  |
| --- | --- | --- |
| **Models** | **Representation** | **Accuracy** |
| ZeroR | Raw | 73.42% |
| MLP | TF-IDF | 90.62% |
| Embeddings | 92% |
| SVM | TFIDF | 91.07% |
| Embeddings | 92.6% |

Table 1. Accuracies of the classifiers on the different feature representations

The SVM model with embeddings representation yielded the highest accuracy of 92.6%.

### **Classification Report**

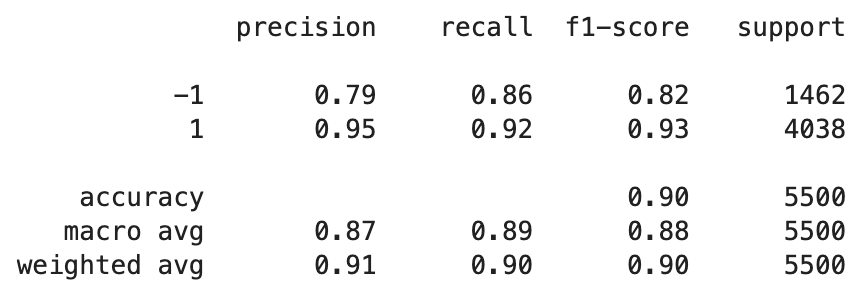


Figure 2. MLP on TF-IDF representation

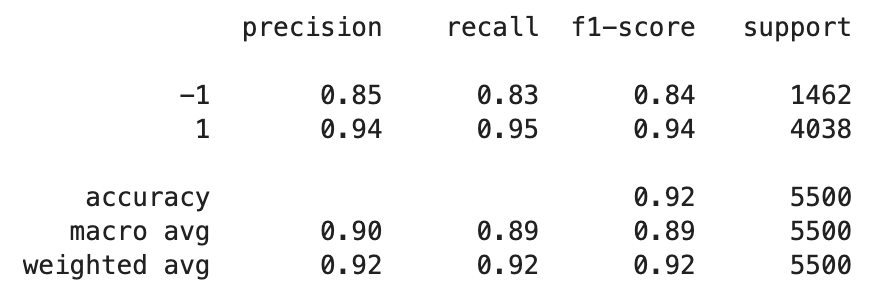


Figure 3. MLP on Embeddings representation

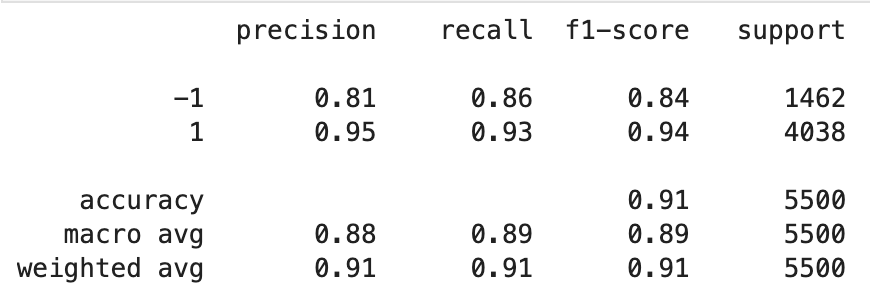


Figure 4. SVM on TF-IDF representation

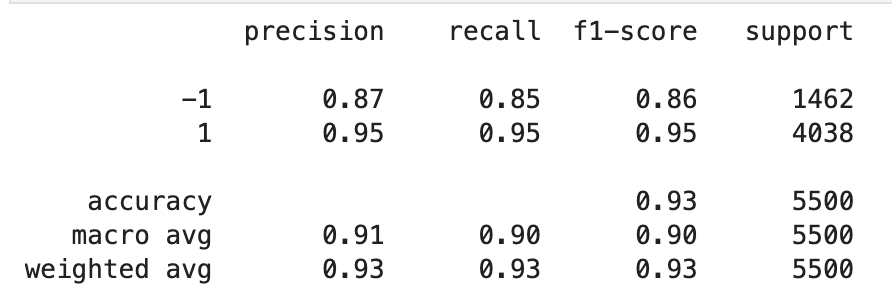


Figure 5. SVM on Embeddings representation

SVM with embeddings outperformed other configurations across all metrics (precision, recall, and F1-score) for both classes. For class -1(negative sentiment), it achieved a precision of 0.87, recall of 0.85, and F1-score of 0.86, while for class 1(positive sentiment), the precision, recall, and F1-score were 0.95.

## **Gender Specific**

### **Initial Observations**

A graph with blue rectangles

Description automatically generated

Figure 6. The distribution of male doctor’s ratings for the training dataset

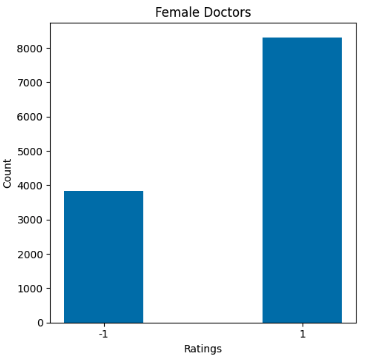


Figure 7. The distribution of female doctor’s ratings for the training dataset

Figures 6 and 7, reveals a distinct difference in the distribution of ratings among male, female clinicians. Female doctors (gender 0) have a total of 3,828 negative and 8,313 positive ratings. In contrast, male doctors (gender 1) have a total of 7,498 negative and 20,013 positive ratings. This representation difference suggests that female doctors receive a higher count of negative ratings in proportion to positive ratings compared to male doctors.

### **Accuracy**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Representation** | **Accuracy** | |
| **Female** | **Male** |
| ZeroR | Raw | 73.81% | 72.5% |
| MLP | Embeddings | 89.2% | 91.9% |
| SVM | Embeddings | 89.7% | 92.5% |

Table 2. Accuracies of the classifiers on the Raw and Embeddings feature representations specific to Male and Female Clinicians.

Both SVM and MLP models exhibit superior performance on male reviews compared to female reviews with the Embedding’s representation.

### **Classification Report**

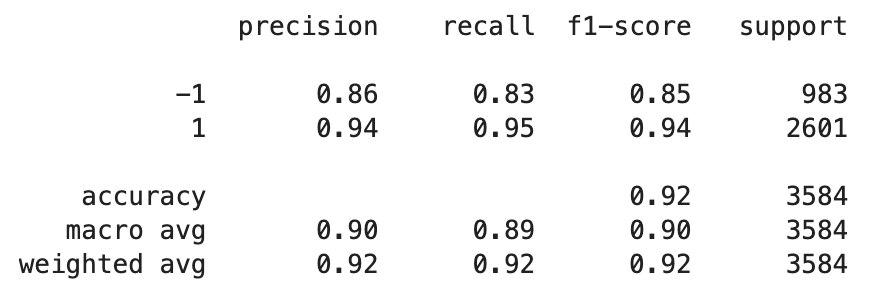


Figure 5. MLP with Embeddings for Male Clinicians

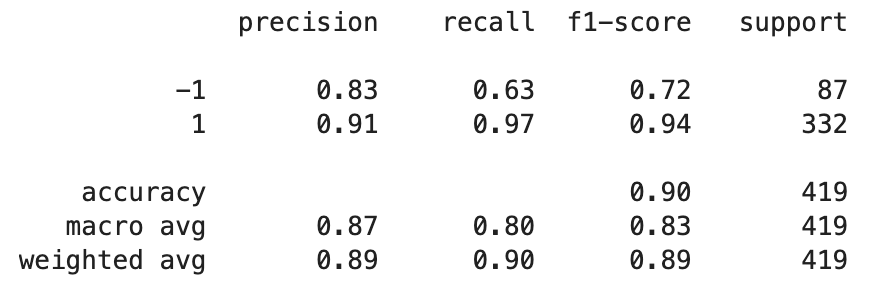


Figure 6. MLP with Embeddings for Female Clinicians

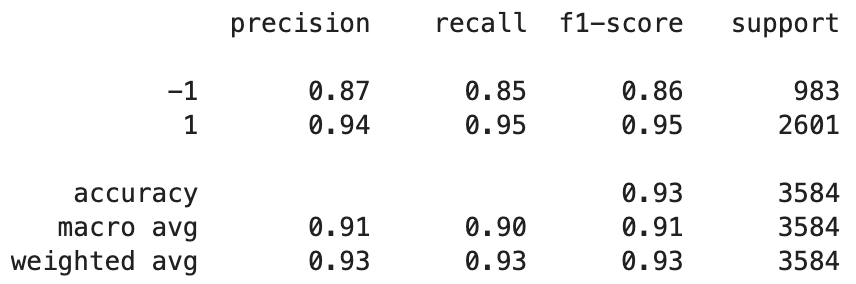


Figure 7. SVM with Embeddings for Male Clinicians

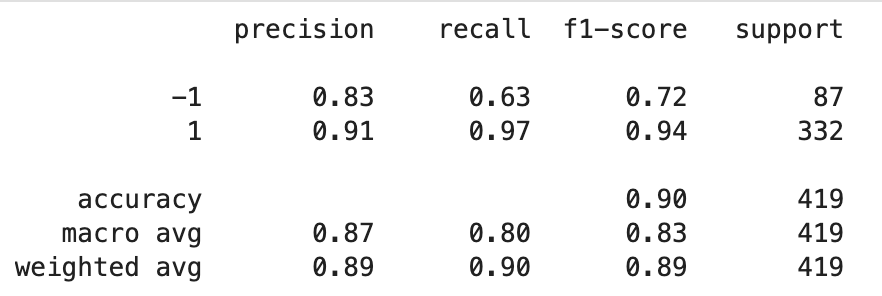


Figure 8. SVM with Embeddings for Female Clinicians

For male reviews, both models exhibited similar performance metrics, with SVM slightly outperforming MLP in terms of precision, recall, and F1-score across both sentiment classes. For female reviews, the performance metrics are identical between SVM and MLP, indicating that the models handled the gender-specific dataset equally well in this smaller sample size.

# **Discussion**

## **Performance Analysis**

Embeddings representation enhanced the performance of both SVM and MLP models compared to TF-IDF representation. This could be attributed to embeddings capturing semantic relationships between words, which is essential in sentiment analysis.

The class imbalance present in the dataset did impact the model's ability to correctly classify sentiments, particularly for the negative class which had fewer instances. This was evident in the lower recall for negative sentiments, indicating that the models had difficulty identifying negative sentiments compared to the positive ones.

Both MLP and SVM managed to handle the class imbalance reasonably well, especially when using the embeddings representation. The F1-score, which is a balanced measure of precision and recall, indicated that the embeddings representation helped in achieving a more balanced performance across both sentiment classes.

Both MLP and SVM showed significant improvements over the baseline model, ZeroR, which had an accuracy of 73.42%. This improvement emphasizes the effectiveness of more complex models and feature representations (embeddings) in capturing the underlying sentiment in reviews.

## **Gender Specific Performance**

The analysis of gender-specific performance revealed a small but apparent discrepancy in accuracy across male and female clinicians. The models appeared to perform better for male clinicians, revealing a potential gender bias.

For female clinicians, the recall for negative sentiments is considerably lower compared to male clinicians. This points to a more profound gender discrepancy in the classification of negative sentiments, suggesting that the models find it more challenging to identify negative sentiments for female clinicians.

Interestingly, the models that excelled on the overall data also exhibited less gender discrepancy, hinting at a more balanced performance across gender groups. This suggests that embeddings might be capturing sentiment in a less biased or more nuanced manner.

The models had more data for male than female clinicians, which might influence the observed performance differences due to sample size imbalance.

Several factors could contribute to the observed gender discrepancy, such as differing vocabulary or expression styles in reviews for male and female clinicians, which might affect the models' ability to accurately classify sentiment.

Despite the gender discrepancy, the SVM model exhibited a marginally better performance across both genders. The consistency in performance across both SVM and MLP models for female clinicians suggests that the observed gender discrepancy is not model-specific but likely related to the dataset.

## **Limitations**

One limitation of the current approach is the reliance on pre-defined feature representations (TF-IDF, embeddings). Future work could explore alternative feature engineering methods to better capture sentiment across different demographics.

There's potential to explore other models and techniques that could further improve performance and reduce gender discrepancy. Additionally, delving into more interpretable models could provide better insights into the sources of bias and how to mitigate them.

# **Conclusions**

In conclusion, this project explored potential gender biases in sentiment analysis using SVM and MLP classifiers. It was clear that the models often favored reviews of male clinicians. However, this bias seemed to stem more from the nature of the dataset than the classifiers themselves. Such biases are especially concerning in fields like healthcare, where fairness is essential.

The project also showed that embeddings are better at representing text than TF-IDF, capturing deeper meanings in the text.

Factors such as distinct review styles or vocabularies for male and female clinicians might contribute to the observed discrepancies. Acknowledging these factors is vital for the development of more equitable machine learning models.

The project highlights the need to be careful about gender biases in machine learning and points out how important good feature representation is. The findings from this research give a starting point for future work that aims to improve sentiment analysis models and make them fair for all genders.

# **Ethics Statement**

The information used in this research must remain confidential, ensuring patients' and clinicians' details are not disclosed due to its sensitive nature. Unauthorized use or access to this data could compromise patient privacy. Patients should also have been informed that their feedback would be utilized in machine learning research. When developing models, it's crucial to reduce biases and ensure predictions are impartial, particularly in healthcare. Transparency in model application is vital, and this report has addressed that. Furthermore, any predictions about clinicians should never be used to harm or negatively affect their well-being.

# **References**

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.

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.

Manning, C. D., Raghavan, P., and Sch ̈utze, H. (2008). Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence

embeddings using Siamese BERT-Networks. In Proceedings of

the 2019 Conference on Empirical Methods in Natural

Language Processing. Association for Computational

Linguistics.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?:

sentiment classification using machine learning techniques. In

Proceedings of the ACL02 conference on Empirical methods in

natural language processing (Vol. 10, pp. 79–86). Association

for Computational Linguistics.

Guia, M., Silva, R. R., & Bernardino, J. (2019).

Comparison of Naïve Bayes, Support Vector

Machine, Decision Trees and Random Forest on

Sentiment Analysis. KDIR, 1, 525-531.

Singla, T., Gaur, V., & Misra, D. K. (2022).

Comparison between Multinomial Naive Bayes and

Multi-Layer Perceptron for Product Review In Real

Time. In 2022 International Conference on Machine

Learning, Big Data, Cloud and Parallel Computing

(COM-IT-CON) (pp. 374-379). Faridabad, India.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016).

Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS) (pp. 4349–4357).

Datta, A., Tschantz, M. C., & Datta, A. (2015).

Automated experiments on ad privacy settings. Proceedings on Privacy Enhancing Technologies, 2015(1), 92–112.

Kiritchenko, S., & Mohammad, S. (2018).

Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (pp. 43-53). https://doi.org/10.18653/v1/S18-2005.