Bias Detection in Physician Notes

CS229 Final Project Milestone

Bikal Sharma
brsharma@stanford.edu

Priyanka Shrestha
shrestp@stanford.edu

1 Motivation

There is observational evidence suggesting biases in clinical documentation [1], yet the development of a deep learning framework to quantify such biases remains unexplored. This project focuses on assessing demographic biases in clinical notes, particularly between genders.[2] We build off of a previous approach which involved employing ClinicalBioBERT, a pre-trained language model specialized for processing electronic health record texts. [3] We hope to extend this project by reproducing the same analysis using large language models like GPT-4. We propose to adapt these LLMs to the specific context of clinical notes. The focus will be on fine-tuning these models for higher sensitivity to subtle language cues that may indicate bias based on gender. This project is application-oriented, aiming to enhance practical tools and methodologies in medical informatics for identifying and mitigating biases in healthcare documentation.

2 Methods

We run experiments based on a comparative analysis of clinical notes with and without demographic identifiers. We first implemented three baseline models: a logistic regression classifier, a support vector machine (SVM) classifier, and a random forest classifier.

Logistic Regression is a linear model used for binary classification tasks that models the probability of the outcome variable belonging to a particular class as a function of the predictor variables.

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n)}} \tag{1}$$

SVM is a supervised learning algorithm used for classification that works by finding the hyperplane that best separates the classes in the feature space.

Equation:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i \quad \text{subject to} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \tag{2}$$

Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputting the mode of the classes.

We used several common evaluation metrics to assess performance. The first metric we used was accuracy: the proportion of correct predictions out of the total number of predictions made.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{3}$$

We also used weighted precision. Precision measures the proportion of true positive predictions out of all positive predictions made by the model, and weighted precision calculates precision for each class and then computes the weighted average based on the number of true instances for each class.

TP_irepresents true positives for class i,
FP_irepresents false positives for class i,
FN_irepresents false negatives for class i,
$N_i$represents total number of classes.

$$\text{Precision} = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N}(TP_i + FP_i)} \tag{4}$$

Next, we use weighted recall (sensitivity). Weighted recall calculates proportion of true positive predictions out of all actual positive instances for each class and then computes the weighted average based on the number of true instances for each class.

$$\text{Recall} = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N}(TP_i + FN_i)} \tag{5}$$

Finally, we use weighted F1 score - the weighted average of the harmonic mean of precision and recall for each class based on the number of true instances for each class.

$$\text{F1-Score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

3 Data and Pre-Processing

We use the fourth edition of the Medical Information Mart for Intensive Care (MIMIC-IV) as our dataset. MIMIC-IV is a collection of deidentified patient electronic health records from Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA. [4] This dataset contains free-text clinical notes in the form of discharge summaries from 145,915 patients pre-processed into single sentences. Each patient with a discharge summary also contains demographic information about sex and race for each patient.

An example discharge note is shown in Figure 1. We use a patient's History of Present Illness section as input for our classification model which outputs a prediction of sex. MIMIC-IV Note data exists as a .csv file where each row contains a clinical note entry. In Figure 2, we show our preprocessing framework for extraction of the input text and output labels from the data, tokenization of sentences, and gender-based terms filtering.

For tokenization, we used scikit-learn's Term Frequency Inverse Document Frequency (TF-IDF) vectorizer. We limited the feature space to the top 3000 terms based on their TF-IDF scores across the corpus to maintain computational efficiency and relevance. Each document was thus represented as a 3000-dimensional vector, where each dimension corresponds to the TF-IDF score of a term in the document relative to its importance across the entire dataset. The choice of TF-IDF vectorization was motivated by its proven effectiveness in enhancing the performance of text classification models by emphasizing words that are critical for distinguishing between document categories.

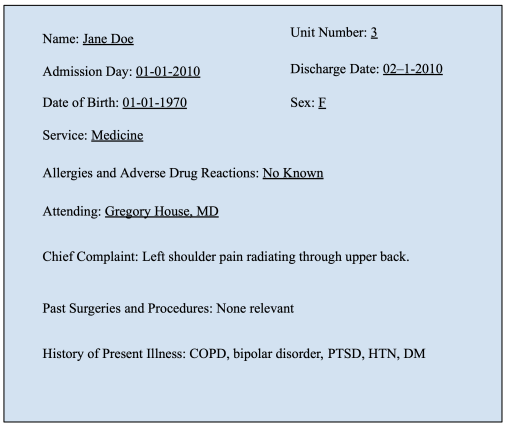

Figure 1: An example of a clinical note from MIMIC-IV. We use 'Sex' as output and 'History of Present Illness' as input.

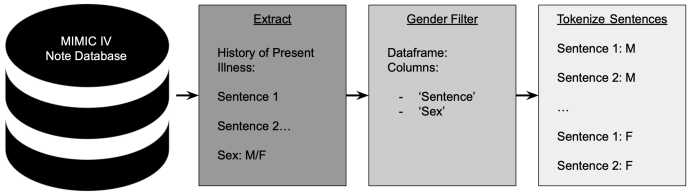

Figure 2: Pre-processing Pipeline

4 Preliminary Experiments

We ran tokenizing on both a filtered dataset (all gender-based words filtered) and an unfiltered dataset and ran the three aforementioned baselines. The results are shown below in Table 1.

Table 1: Baseline Model Evaluations

Model	Accuracy	Precision	Recall	F1
logistic_regression_f	0.53	0.53	0.53	0.53
svm_f	0.525	0.521	0.525	0.459
random_forest_f	0.528	0.525	0.528	0.524
logistic_regression_uf	513	0.51	0.512	0.51
svm_uf	0.52	0.51	0.52	0.44
random_forest_uf	0.517	0.516	0.517	0.516

Model names ending in '_f' are ones trained on data with gendered words filtered whereas model names ending in '_uf' are trained on data without filtering.

Overall, our baseline performed rather poorly with accuracy metrics only just above 50%. Our best performance based on accuracy was by the random forest classifier on filtered data with an accuracy of 0.528. In fact, on average the baseline models run on the unfiltered dataset were worse than the baseline models run on the filtered dataset. This is the opposite of what we expected, because intuitively having gendered-terms like pronouns should make it easier classifications models to differentiate between notes for males versus females. This most likely points to issues with the way we have tokenized our sentences, leading to inappropriate representations that are not conducive to accurate classification.

5 Next Steps

We plan to do the same experiments done on the baseline models on deep learning models. In addition to improve from the baseline model embeddings, we will add a set of experiments utilizing contrastive learning (such as SimSCE) to enhance embeddings and subsequent classification. [5]

We will use a sequence classification mode with both ClinicalBioBERT and an LLM, predicting if the given physician note is in reference to a particular gender. We will compare models with and without masking of any gender identifiers and evaluate using a fill-in-the-blank approach to assess the neutrality of language and the presence of implicit biases. We will use disparity metrics to evaluate the differences in language use across genders, including differential language usage rates and cluster distribution disparities in gendered terms

There are many possible future research directions in the realm of this project. We specifically investigated gender biases, but these analyses could be applied to other demographic groups such as race, ethnic identity, class, LGBTQ identity, etc. to allow for a more comprehensive understanding of biases in clinical documentation. In this same vein, one might also pursue multilabel classification int order to understand intersectional biases in clinical notes. [6] For these projects, as well as any seeking to understand gender biases better, one might also improve sentence embeddings by using tools like masking multitoken words, or using a large masking vocabulary, as well as better filtering tools to try to extract clinician comments specifically on the patient and their presentation as it relates to the demographic variables we are interested in. Finally, we are interested in generalizing our analyses to other hospital data sets, as well as specialty and disease specific datasets in order to see the spectrum of gender biases in different types of medical institutions and fields.

6 Contributions

BS wrote motivations, methods, data and preprocessing, half of next steps, and made figures. PS wrote preliminary experiments, half of next steps, and made tables. BS extracted and preprocessed data, PS ran and evaluated baseline models.

References

[1] Gracie Himmelstein, David Bates, and Li Zhou. "Examination of Stigmatizing Language in the Electronic Health Record". In: *JAMA Network Open* 5.1 (Jan. 2022), e2144967. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2021.44967. URL: https://doi.org/10.1001/jamanetworkopen.2021.44967 (visited on 02/20/2024).

[2] Anthony Rios, Reenam Joshi, and Hejin Shin. "Quantifying 60 Years of Gender Bias in Biomedical Research with Word Embeddings". In: *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. Ed. by Dina Demner-Fushman et al. Online: Association for Computational Linguistics, July 2020, pp. 1–13. DOI: 10.18653/v1/2020.bionlp-1.1. URL: https://aclanthology.org/2020.bionlp-1.1 (visited on 02/20/2024).

[3] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805 [cs]. May 2019. DOI: 10.48550/arXiv.1810.04805. URL: http://arxiv.org/abs/1810.04805 (visited on 02/20/2024).

[4] Alistair E. W. Johnson et al. "MIMIC-IV, a freely accessible electronic health record dataset". en. In: *Scientific Data* 10.1 (Jan. 2023). Number: 1 Publisher: Nature Publishing Group, p. 1. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01899-x. URL: https://www.nature.com/articles/s41597-022-01899-x (visited on 02/20/2024).

[5] Tianyu Gao, Xingcheng Yao, and Danqi Chen. *SimCSE: Simple Contrastive Learning of Sentence Embeddings*. arXiv:2104.08821 [cs]. May 2022. DOI: 10.48550/arXiv.2104.08821. URL: http://arxiv.org/abs/2104.08821 (visited on 02/20/2024).

[6] David M Markowitz. "Gender and ethnicity bias in medicine: a text analysis of 1.8 million critical care records". In: *PNAS Nexus* 1.4 (Sept. 2022), pgac157. ISSN: 2752-6542. DOI: 10.1093/pnasnexus/pgac157. URL: https://doi.org/10.1093/pnasnexus/pgac157 (visited on 02/20/2024).