Bias in clinical notes

Stanford CS224N Custom Project

Betty Xiong
Department of Computer Science
Stanford University
xiongb@stanford.edu

Abstract

There has been anecdotal evidence of bias in clinical notes, and previous studies have attempted to quantify the differences between different groups of patient attributes, e.g. female vs. male, Black vs. non-Black. However, there has yet to be a deep learning model that quantifies this bias. This project aims to to measure demographic bias between female and males in clinical notes, whilst accounting for confounding factors such as gender- and disease-specific words. The model utilizes a pretrained language model, ClinicalBioBERT that is trained specifically on electronic health record corpora, and trains sentences labelled with female/male to pass through a classification layer. An extra qualitative analysis with TF-IDF and K-means clustering demonstrates evidence of bias in different words used for male vs. female patients. The sequence classification model utilizes masking tokens to mask away confounding factors such as gendered and disease specific words - this is compared against a baseline.

1 Key Information to include

- Mentor: Ansh Khurana
- External Collaborators (if you have any): Dr. Barbara Engelhardt
- Sharing project: NA

2 Introduction

Bias in the medical field has been a topic of interest. Rios et al. (2020) quantified 60 years of gender bias in biomedical research with skip-gram model of word embeddings. In the clinical context, studies show gender and ethnicity bias, with stigmatizing language appear in the admission notes and critical care records of patients who are hospitalized (Markowitz, 2022; Himmelstein and Zhou, 2022). There has been research that examines the extent to which embeddings may encode marginalized population differently, and how this may lead to a perpetuation of biases and worsened performance on clinical tasks (Zhang et al., 2020). The paper demonstrates dangerous latent relationships from contextual word embeddings, and performance gaps across different definitions of fairness on downstream clinical prediction tasks.

Despite previous research, there has not been any approach on deep learning transformer models. We aim to to measure demographic bias between female and males in clinical notes. The model utilizes the clinical medical domain-specific pretrained language model ClinicalBioBERT to learn sentence-level representations and generates the predicted tokens with a transformer decoder. For data, our model uses free-text notes from the database MIMIC-IV-Note: De-identified free-text clinical notes (2008-2019) (Johnson et al., 2023). Specifically, using the patient's 'History of present illness' and 'Sex' sections of the clinical notes, we assign them as 'text' and 'label' respectively, for a text classification task, analagous to sentiment analysis. The model aims to predict a test set of notes and classify them as either for a 'male' or 'female' patient.

Due to confounding factors in predicting male or female attributes in each clinical note, our model introduces a masking technique to insert mask tokens to disregard i. gender-specific words and pronouns, e.g. she/he, female/male; ii. disease-specific terms, i.e. disease words collected from Medical Subject Headings (MeSH) (NLM, 2023) thesaurus is a controlled and hierarchically-organized vocabulary produced by the National Library of Medicine. It is used for indexing, cataloging, and searching of biomedical and health-related information.

In this project, we aim to to measure demographic bias between female and males in clinical notes, whilst accounting for confounding factors such as gender- and disease-specific words.

3 Related Work

Tokenizer. We use a standard tokenizer from BERT (Bidirectional Encoder Representations from Transformers) models rely on a word-slice tokenizer that has been trained on Wikipedia and literature datasets. From that, we use the idea of incorporate masking tokens for gender- and disease- specific (MeSH) (NLM, 2023) before passing the input into the model.

Pretrained language model. Pretrained language models are large neural networks that are used in a wide variety of natural language processing (NLP) tasks. They operate under a pretrain-finetune paradigm: Models are first pretrained over a large text corpus and then finetuned on a downstream task. The most commonly used ones such as BERT (Devlin et al., 2019) and DistilBERT (Sanh et al., 2020) have not been pretrained on specialty corpora such as clinical text. Without the domain-specific text, performance is poor. BioBERT (Devlin et al., 2019) is a BERT model finetuned on PubMed abstracts and PMC articles. ClinicalBioBERT is a model that further fine-tunes on clinical texts from about two millions notes of all types in the MIMIC-III (Johnson et al., 2016) to make ClinicalBioBERT, which fits the semantics of clinical notes well.

Classification. Classification based on sentiment analysis (Dang et al., 2020) has been used in a variety of contexts, but not in bias for the medical domain. Other methods for defining classification problems involves contrastive learning, whose main idea is to learn effective representation by pulling semantically close neighbors together and pushing apart non-neighbors. Gao et al. (2021) presents an effective contrastive learning framework for sentence embedding, called SimCSE, which can be used downstream for classification tasks.

4 Approach

Main approach. The approach is to use a sequence classification mode . The architecture is as follows (Figure 1): we insert a [CLS] token at the beginning of each sentence, and takes a single sentence and to generate token embeddings. We integrate a second step of introducing token masks for vocabulary that pertains to diseases, as per MeSH (NLM, 2023). From the MeSH descriptor vocabulary, we take terms from the top-level categories in the hierarchy: Anatomy [A] and diseases [C], to create a custom masking function within the tokenizer framework. The baseline experiment does not include the masking function, whereas the main experiment does. This is fed into the ClinicalBioBERT Transformer encoder, and the classification performed after passing it through a GELU (Gaussian Error Linear Unit) layer.

K-means clustering. To further identify potential terms that are salient to particular documents or document classes, we apply term frequency–inverse document frequency (TF-IDF) as a weighting factor to the input data. TF-IDF is a numerical static that evaluates how relevant a word is to a document in a collection of documents or corpus. This is achieved by combining two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents. This allows us to identify common words within specific notes, that are unique to a specific gender. From that, we use K-means clustering algorithm (k=8) using TF-IDF weighted word embeddings to analyze any patterns discovered amongst articles within the different genders.

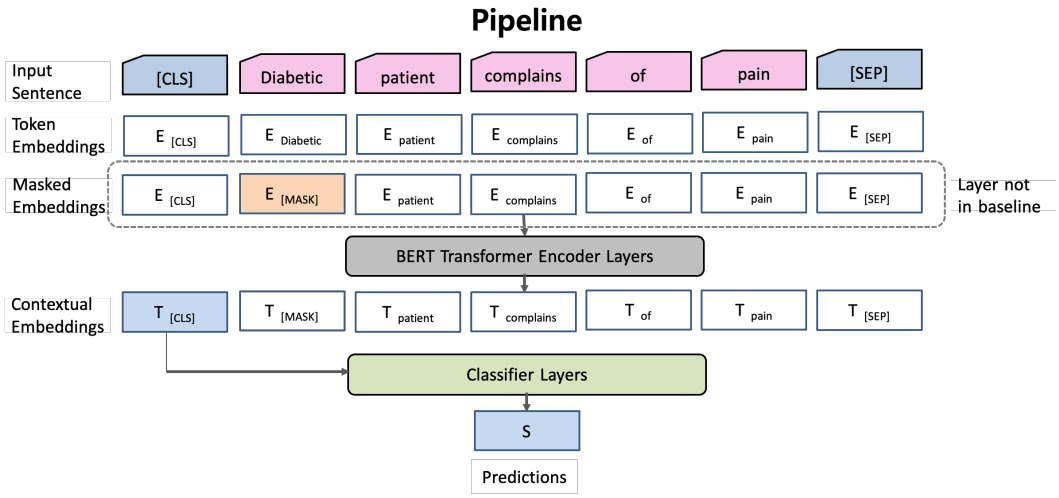

Figure 1: Architecture of neural network. It takes single sentences and generates token embeddings. A second (optional, baseline does not include it) step of masks disease-related words before passing it into the ClinicalBioBERT Transformer encoder. Only [CLS] token embeddings are used as inputs for the following layers with the extractive objective for predicted classification for each sentence.

	Train	Val	Test
Female	6499	1594	2554
Male	6301	1606	2446

Table 1: Data split

5 Experiments

5.1 Data

Data source. MIMIC-IV is a publically available database sourced from the electronic health record of the Beth Israel Deaconess Medical Center. The specific dataset used is MIMIC-IV-Note: De-identified free-text clinical notes (2008-2019) (Johnson et al., 2023), a collection of deidentified free-text clinical notes for patients included in the MIMIC-IV clinical database. MIMIC-IV-Note contains 331,794 deidentified discharge summaries from 145,915 patients admitted to the hospital and emergency department. All notes have had protected health information removed in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor provision.

Data format and preprocessing. The data was in the form of a .csv file, where each row is a note entry. The section 'History of present illness' is used as input for the classification model, and 'Sex' is the output prediction. We show an example discharge note in Figure 2. The data is split as per Table 1. There were some preprocessing steps before entering the model, as outlined in Figure 3 They include: extracting the appropriate input (text) and output (label) from the file, tokenizing the notes into separate sentences, and filtering any sentences that include clinical observation, e.g. In the ED, initial vitals were 98.9 88 116/88 18 97% RA. CBC near baseline, INR 1.4, Na 125, Cr 0.6. AST and ALT mildly above baseline 182 and 126 and albumin 2.8.

5.2 Evaluation method

Evaluation Metrics. Standard text classification evaluation techniques include F1 score, accuracy, precision and recall. Accuracy measures how many observations, both positive and negative, were correctly classified. Precision measures how many observations predicted as positive are in fact positive. Recall measures measures how many observations out of all positive observations have we classified as positive. F1 is the harmonic mean between precision and recall.

Name: ___ Unit No: ___

Admission Date: ___ Discharge Date: ___

Date of Birth: ___ Sex: F

Service: MEDICINE

Allergies:
No Known Allergies / Adverse Drug Reactions

Attending: ___

Chief Complaint:
Worsening ABD distension and pain

Major Surgical or Invasive Procedure:
Paracentesis

History of Present Illness:
___ HCV cirrhosis c/b ascites, hiv on ART, h/o IVDU, COPD,
bioplar, PTSD, presented from OSH ED with worsening abd distension over past week.

Figure 2: An example data record from MIMIC-IV-NOTE (Johnson et al., 2023) We use 'History of present illness' and input and 'Sex' as output. Note that all patient sensitive information has been replaced with "___".

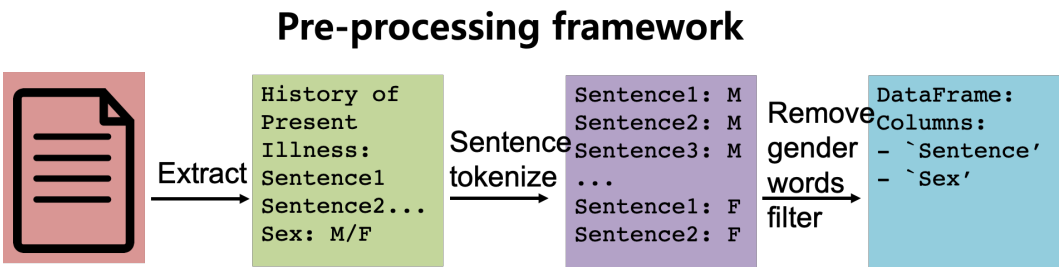

Figure 3: Pre-processing pipeline for extracting sentences from clinical notes

5.3 Experimental details

Hyperparameters for ClinicalBioBERT Classifier:
Training time = 8 hours 16 minutes on AWS VM with 32 GB memory
num train epochs=2
learning rate=2e-5, lr scheduling=linear warmup, linear decrease
optimizer=AdamW, batch size=64, early stopping=True
dropout=0.1, hidden dim=3072, max position embeddings=512
nheads=12, nlayers=6

5.4 Results

The results show that the model classifies slightly above random. Table 2 shows the training loss and validation accuracy, whereas Table 3 shows the accuracy, F1 score, precision and recall of the test set. In general, the baseline and disease masking perform roughly the same, which is surprising, because one would assume that the masked model would perform worse, given there is less information to be gleaned from the disease input terms. However, this may show that the diseases are either not all completely masked, or that they do not have significant contribution to the classification task. For both the baseline and disease masking models, the recall (about 0.68) is better than the recision (about 0.53).

Model	Training Loss	Accuracy
Baseline	0.6914	0.525
Disease masking	0.6906	0.5244

Table 2: Metrics for baseline and with disease masking (evaluation set)

Model	F1	Accuracy	Precision	Recall
Baseline	0.5242	0.6019	0.5349	0.6812
Disease masking	0.5224	0.5917	0.5252	0.6774

Table 3: Metrics for baseline and with disease masking (test set)

6 Analysis

Model classifies slightly above random. There are several reasons for the failure modes.

Data. The data is not curated well, i.e. there is not good quality data that pertains to a clinician's qualitative and commentary on the patient, i.e. data is like a doctor's scratchpad and needs to be further cleaned up., masking is too primitive. As well, we are limited by the informativeness of the data. We only have labels for the patient gender, but not that of the clinician. Since communication is a two-way exchange, without the information of the clinician gender, our analysis is limited. As well, the limited compute meant that not all of the data could be processed.

Clustering analysis. TF-IDF for K-means clustering, to find terms that are most informative for each cluster. Figure 4 demonstrates the cluster projected onto a 2D space. The number of clusters chosen was eight, due to the lowest sum of square errors. We allow for more clusters than two (pertaining to male and female genders) because there may be more attributes that are not completely captured within two clusters. For one of the clusters where the majority of data points belonged to a female patient, the top key words were:

`floor, complain, palpitations, denies, back, report, chest, abdominal, pain,`

as opposed to a predominately male assigned cluster, which included top key words as:

`fat, provide, content, shake, inform.`

The clustering analysis suggests that there are differences in language used on patients of different genders, however, the initial model was not able to classify those differences. The limiting factors have been highlighted above.

7 Conclusion

The main outcome of this project was a successfully developed a pipeline for classifying a sentence in a clinical note from electronic health records as more likely to be from a female or male. We also deployed an initial version of clustering sentences to allow for further qualitative analysis of possible failure modes.

Future directions. For future research, we can improve sentences embeddings, e.g. use contrastive learning (such as SimSCE) to enhance embeddings and subsequent classification. We could then use a loss function that computes word similarity as the distances in the embedding space. We can have better masking, by removing confounding input. This can be achieved by masking multi-token words (as disease specific words tend to be longer, and sentiment specific words are simpler and shorter), or use larger vocabulary of terms for masking. We can deploy better filtering to only include sentences where the clinicians comment on how the patient presents themselves. Finally, an interesting subproblem would be to have disease-specific datasets, e.g. a disease that affects men and women equally like cardiovasclar diseaese or diabetes, such that the question of disease-related words is negated.

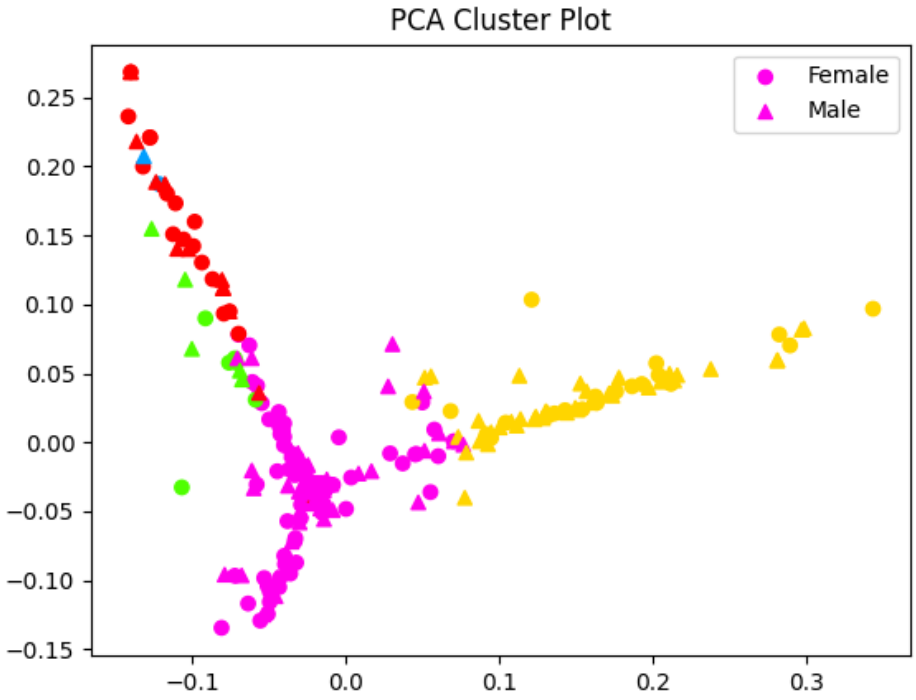

Figure 4: PCA plot of K-means (n=8) clustering after TF-IDF analysis.

References

Nhan Cach Dang, María N. Moreno-García, and Fernando De la Prieta. 2020. Sentiment analysis based on deep learning: A comparative study. In *arXiv data*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *ACL Anthology*. Association for Computational Linguistics.

Tianyu Gao, Xingchen Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Association for Computational Linguistics (ACL)*.

Gracie Himmelstein and David Batesand Li Zhou. 2022. Examination of stigmatizing language in the electronic health record. In *JAMA Netw Open*. Journal of the American Medical Association.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. Mimic-iv-note: Deidentified free-text clinical notes. In *PhysioNet*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. In *Scientific data*. Nature.

David M Markowitz. 2022. Gender and ethnicity bias in medicine: a text analysis of 1.8 million critical care records. In *PNAS Nexus*. The National Academy of the Sciences of the United States of America.

NLM. 2023. Medical subject headings (mesh). In *National Library of Medicine*.

Anthony Rios, Reenam Joshi, and Hejin Shin. 2020. Quantifying 60 years of gender bias in biomedical research with word embeddings. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 1–13. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *arXiv*.

Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical embeddings. In *CHIL '20: Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120. Association for Computating Machinery.