Investigating BioBERT Performance on Gender Bias

Dayoung Kim, MS, Erin Zeng, MS, Nikolay Lukyanchikov, MS

Weill Cornell Medicine, New York, United States

April 10, 2022

Abstract

Natural Language Processing (NLP) is an approach commonly used in biomedical research to analyze unstructured clinical data. Historical studies suggest that NLP models can often reproduce and amplify the bias from the input data and proliferate this bias in downstream tasks. In this study, we performed an experiment using BioBERT model, which is a transformer-based model pre-trained on PubMed data, to perform a classification task of labeling gender based on clinical trial notes. The model produced an Accuracy score of 0.7143 and Precision score of 0.3571 ('Male' = 0, 'Female' = 1), indicating the model was biased towards male. We discussed the significance of these results and proposed directions for future studies.

1.0 Project Definition

Natural Language Processing (NLP) is commonly used in biomedical research to analyze unstructured clinical data. Despite NLP performance being relatively promising, research suggests that the results could be subject to several types of biases, such as demographic bias, linguistic bias, and redundancy bias. These biases post a challenge in interpreting the results correctly as the outcome could be heavily skewed to a dominant group. Although the concept of bias is still relatively underexplored in the NLP community, addressing this challenge is important because biases can proliferate downstream and affect further interpretation of results by clinicians and biomedical researchers.

Bidirectional Encoder Representations from Transformers (BERT) is one of the most powerful NLP models in recent years. Due to its transformer-based architecture, BERT offers enhanced parallelization and better modeling of long-range dependencies (1). It is much faster and more space-efficient than non-transformer models.

BioBERT, or Bidirectional Encoder Representations from Transformers for Biomedical Text Mining, is a domain-specific language representation model pre-trained on large-scale biomedical corpora (2). BioBERT has been shown to outperform base-cased BERT models on biomedical text mining tasks including biomedical named entity recognition, biomedical relation extraction, and biomedical question answering (2).

In this study, we will use a pre-trained BioBERT model to perform a text classification task of gender based on clinical notes detailing patient conditions and activities from a clinical trial. We will evaluate the performance results of the model and conclude whether BioBERT is subject to gender bias. The results will be summarized and the models will be shared with the clinical community to provide guidance on future research.

2.0 Related work

NLP bias is a fairly new and underexplored area within NLP research. A literature review conducted by Straw et. al in 2020 looked into 52 papers to see where the authors had investigated biases within their model but found only two papers of 52 had assessed such biases within the NLP model architecture (3). In her study, Straw chose to examine how GloVe and Word2Vec, two of the most popular word embedding models, performed on bias against mental health data in a social media context. The results showed that both GloVe and Word2Vec embeddings demonstrated significant biases with respect to religion, race, gender, nationality, sexuality and age. For example, the GloVe model attributed male related words to "violent" and female related words to "innocent" (3).

Bhardwaj et al. trained a simple regressor utilizing BERT's word embeddings on 5 downstream tasks related to emotion and sentiment intensity prediction to investigate gender bias in BERT. The regressor they built was a shallow multi-layer perceptron (MLP) without fine-tuning BERT parameters, which they believed would better expose inherent gender bias in BERT. The tasks were further classified into emotion intensity regression and sentiment intensity regression, where the regressor was to determine an intensity value given a tweet. Results suggested that the regressors consistently assigned high values to either of the genders and rarely assigned equal scores to both the genders, indicating the model has a significant dependence on gender-particular words and phrases (4).

In another study, word probabilities taken from the BERT language model were used to calculate association bias between a gender denoting target word and an attribute word, such as a profession. Results showed that male person words were relatively stable in BERT. Associations were less affected by the presence of the profession words, and also less affected by fine-tuning. The researchers concluded that this indicates a strong male bias in BERT (5).

Other researchers have explored methods to reduce the gender bias introduced by NLP models. Bolukbasi et al. provided a methodology for modifying an embedding to remove gender stereotypes, while maintaining desired associations between the words. They defined metrics to quantify both direct and indirect gender biases in embeddings, and developed algorithms to "debias" the embedding. The results proved to be usable in applications without amplifying gender bias (6).

3.0 Method

3.1 Baseline

To establish the baseline for our study, we found a research (7) where an evaluation was performed on multiple variations of BERT models' performance on text mining tasks across different data sets in the biomedical domain. Precision, Recall, F1-Score and Accuracy were calculated for each of the models tested. BioBERT was one of those models. We gathered the performance scores of BioBERT, presented in **Table 1**, and chose the Precision and Accuracy values as the baseline. Overall, BioBERT demonstrates a high performance on biomedical text mining tasks, with an average Precision of 0.81 and average Accuracy of 0.85.

Table 1. BioBERT performance on text mining tasks across 11 different datasets in the Biomedical domain

	NCBI[1]	BC5CDR[2]	Species[3]	JNLPBA[4]	GAD[5]
Precision	0.88	0.86	0.73		0.77
Recall	0.91	0.88	0.75		0.83
F1-Score	0.90	0.87	0.74		0.80
Accuracy		0.93		0.78	

[1] NCBI: National Center for Biotechnology Information Disease Corpus (7)
[2] BC5CDR consists of chemical induced disease (CID) relation extractions (7)
[3] MedNLI consists of medical history of the patients which is annotated by doctors (7)
[4] JNLPA: Joint Workshop on Natural Language Processing in Biomedicine and its Application is a corpus of Pubmed abstracts specialized for NER tasks (7)
[5] GAD: Genetic Association Database. Generated from the Genetic Association Archive (7)

Table 1. BioBERT performance on text mining tasks across 11 different datasets in the Biomedical domain (Cont.)

	EUADR[6]	CHEMPROT[7]	MedSTS[8]	Biosses[9]	MedLNI	HOC[10]
Precision	0.85	0.77				
Recall	0.91	0.76				
F1-Score	0.81	0.76				
Accuracy			0.85	0.83	0.81	0.83

3.2 Our System

3.2.1 Data Source

The Harvard DBMI Data Portal provides a wide range of datasets for Biomedical Informatics research. The n2c2 NLP Research Data Sets contains unstructured notes from the Research Patient Data Registry at Partners Healthcare. Specifically, we requested the 2018 (Track 1) - Clinical Trial Cohort Selection dataset (8) and was granted access. We obtained around 300 XML files of clinical trial notes for our experiment.

3.2.2 Pre-processing

Since the files were in XML format, there were a lot of tabs and multiple spaces in the data. We used regular expression to replace multiple tabs and spaces with a single whitespace, so they would not be considered as multiple tokens.

Because this is a supervised learning task, we first need to label the data. We used regular expression to find in all sentences that contain gender-specific words such as ‘She’, ‘He’, ‘Male’ and ‘Female’ and labeled it as 0 - Male and 1 - Female for each sentence. We validated the labeling and confirmed it was correct.

After pre-processing data with labeling, data was split into training dataset and validation dataset using a ratio of 0.7. All of the test dataset were used. Results were stored as TXT files. **Figure 2** and **Figure 3** are examples of the results. After that, we used word piece tokenizer to convert text to token pieces. Final inputs to our model include token and segment.

3.2.3 Fine-tuning

A pre-trained BioBERT model was used for the experiment. This model was pre-trained on over 1M records of PubMed data and was trained in the same way as the BERT base-cased model.

We performed fine-tuning of the model using our dataset. There are two inputs of the model, token and segment. Since there are only 2 labels, we did not use the ReLU function. A sigmoid function was applied to the output layer

---

[6] EUADR: a corpus that contains annotations of multiple entities (drugs, diseases, and targets) and relationships between these entities (15)
[7] CHEMPROT: chemical protein interaction corpus generated from PubMed abstracts (7)
[8] MedSTS: Semantic Textual Similarity in Medical Domain. Consists of a total of 174,629 sentence pairs gathered from a clinical corpus at Mayo Clinic (16)
[9] Biosses: Biomedical Semantic Similarity Estimation System
[10] HOC: Hallmarks of Cancer dataset. Has binary labels which focuses on labeling the cancer discussions on the abstracts as positive samples (7)

to convert the output into a probability score between 0 to 1. And we made the final model take the gender specification layer as a final layer. Last but not least, we compiled the model before we began training the model.

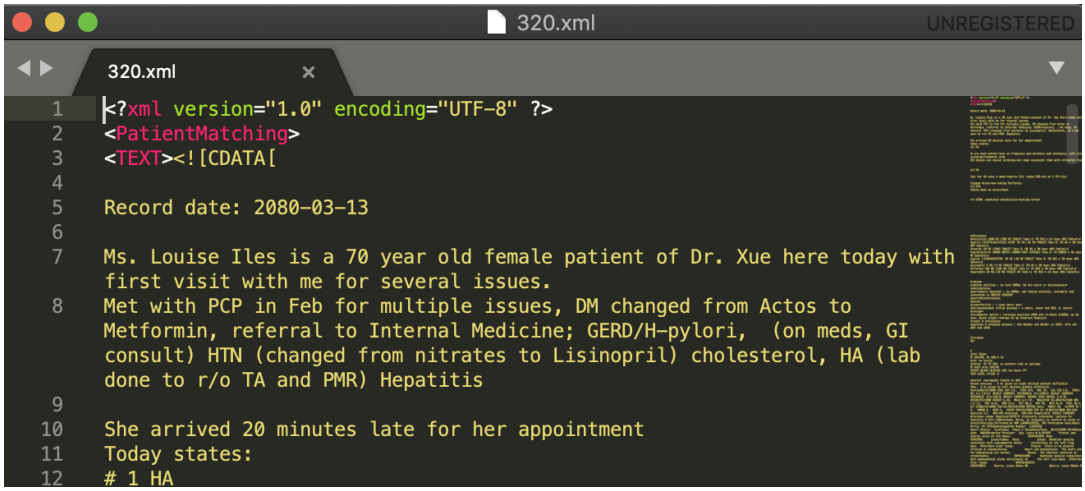
4.0 Experiments

4.1 Experimental Settings

To run the BioBERT model, we used tensorflow and keras. Adam from keras was used as an optimizer. Pandas and Numpy were used to handle datasets and convert them into input format of the BioBERT model. To retrieve sentence format from raw data and replace multiple whitespaces to a whitespace, regular expression was needed with re package. Sklearn was used to get the performance scores by comparing the predictions with the actual labels. To deal with xml format datasets, xml.etree.ElementTree was needed to read the files.

Harvard n2c2-NLP-Research-Data-Sets/2018 (Track 1) - Clinical Trial Cohort Selection was the datasets we used (8). Train file and test file exist with XML files as **Figure 1**. Clinician notes were under <TEXT> tag with multiple '\n's as splitting rows. To get rid of multiple '\n's and words that are not sentences, such as 'Record date: 2080-03-13', regular expression was needed. To replace the blanks to a white space, "re.sub(' [\n]{2,}',' ',description.text)" was used (description.text is the text under <TEXT> tag). To only retrieve sentences," re.findall('[0-9]*[A-Z][^.!?]+[.!?][\n]',temp)" was used (temp is the output of "re.sub(' [\n]{2,}',' ',description.text").

Figure 1. Example of XML File



For the BioBERT model, labeled data is needed. Due to the clinician notes, lots of gender specifying words, such as 'she/he', 'her/his', and 'female/male', were used. For example, 'She arrived 20 minutes late for her appointment' and 'Ms. Louise Iles is a female patient'. Therefore, for each sentence, if it contained such words, it was labeled with 0-male and 1-female. If the sentence doesn't contain gender-specifying words, it was considered that it's from the same report as the previous sentence. After preprocessing data with labeling, training dataset and validation dataset were splitted with the ratio of 0.7. All of the test dataset were used. **Figure 2** and **Figure 3** are examples of the results.

Figure 2. Training Set TXT File

train.txt

```
0   1
"He is a set designer at Columbia Pictures.
"   0
"Diagnosis:  Left ankle fracture.
"   0
"This is a brief addendum to the medical record.
"   0
"The patient has no chest pain at the time of evaluation in the emergency department and no shortness of breath.
"   0
```

Figure 3. Validation Set TXT File

valid.txt

```
0   1
"He was seen by psychiatry here.
"   1
"Continue depakote and lithium (had subthereapeutic level that psych thought was likely  due to non-compliance.
"   1
"He has remained on NPH with RISS.
"   0
"Cath c/b VF arrest after dye load and resultant afib with RVR.
"   0
```

To use the BioBERT model, word piece tokenizer was used. It allows words with similar meaning to have similar representation. For example, if the input text is " the man jumped up, put his basket on philammon's head", the tokenizer would split it into tokens as ["the", "man", "jump", "##ed", "up", "put", "his", "basket", "on", "phil", "##am", "##mon","", "s", "head"]. For tokens, vocabulary from the pretrained BioBERT model would be used.

BioBERT v1.1 was used. It's based on the BERT-base-Cased model with 1 million of PubMed data trained. To use this model, tokens were converted to indices with the model's vocabulary and segment ids were added as input for segmenting sentences. As a result, the dataset's shapes were as **Figure 4**. After loading the pretrained model, to get gender specification output, it was required to finetune the model. By using "keras.layers.Dense(1,activation = 'sigmoid')(pooled_output)", the fine-tuned model takes the output of the pretrained BioBERT model as an input. Then, "keras.models.Model(inputs, output_layer)" was added to get the predictions on gender that are needed. The final model has 12 layers of encoders, which are from the pretrained BioBERT model, dropout layer, lambda layer, and 2 dense layers from fine-tuning.

Figure 4. Shape of Datasets

```
train dataset's shape
(2, 3959, 128)
(1, 3959, 1)
---------------
valid dataset' shape
(2, 1697, 128)
(1, 1697, 1)
---------------
test dataset' shape
(2, 2408, 128)
(1, 2408, 1)
```

Due to previous research, it is commonly known that the optimal values of maximum sequence length, epochs, and learning rate are 128, 3, le-5. The batch size was set as 16. After training the model, the test dataset was used to get predictions. The prediction values are between 0 and 1. The values were rounded, considering the values no less

than 0.5 represent female specifications. With the sklearn package, the performance scores were calculated by comparing the prediction values with the actual values.

4.2 Quantitative Results

4.2.1 Comparison with Baseline

The final model produced an Accuracy score of 0.7143 and Precision score of 0.3571. The Accuracy score suggests that 71.43% of predictions the model made were correct. On the surface this is an acceptable score. However, when compared with the average Accuracy score of 0.85 BioBERT achieves on general biomedical text mining tasks. The score appears to be much lower than our expectation.

On the other hand, while the average Precision score BioBERT model produces on text mining is 0. 81, our model only reached a precision score of 0.3571, indicating only 35.71% of female predictions were correct, as we label 'Female' as 1.

Recall and F1-Score were also calculated for the model. Recall was 0.5000 and F1-Score was 0.4167. However, we chose to evaluate the performance against the baseline model based on Accuracy and Precision only because we were interested in learning how accurately our model made predictions of male vs female, and Accuracy and Precision were more relevant to our research question.

Compared with our baseline performance metrics, we conclude that the BioBERT model makes better predictions on male than female. Results demonstrate that BioBERT is subject to gender bias, as proven in previous studies on BERT and other NLP models

4.2.2 Comparison of the Models with Different Parameters

During our experiment, we tried different parameters to check if they affect performance scores. As a result, the accuracy differed but not significantly. Our original batch size is 16, epochs is 3, and sequence length is 128. **Table 2** shows the comparison among parameters.

Table 2. Performance Metrics with Different Parameters

	Accuracy	Precision	Recall	F1-Score
BATCH_SIZE				
2	0.6691	0.3571	0.5000	0.4167
6	0.7004	0.3571	0.5000	0.4167
8	0.7013	0.3571	0.5000	0.4167
16	0.7143	0.3571	0.5000	0.4167
SEQ_LEN				
32	0.7017	0.3571	0.5000	0.4167
128	0.7143	0.3571	0.5000	0.4167

256	0.6881	0.3571	0.5000	0.4167
do_lower_case				
yes	0.7143	0.3571	0.5000	0.4167
no	0.6992	0.3571	0.5000	0.4167
EPOCHS				
3	0.7143	0.3571	0.5000	0.4167
5	0.7130	0.3571	0.5000	0.4167
10	0.7130	0.3571	0.5000	0.4167

As we can see, there is no significant difference between different parameters, so we decided to use optimal parameters mentioned in the Experiment settings section.

4.3 Discussion Qualitative

It is shown that the BioBERT model can be used to deal with clinician notes, which are unstructured data. However, even though the accuracy was relatively high, the precision score was low. Since the notes about female patients were labeled as '1', the predictions made as female patients weren't correct. Considering the previous research and the performance scores of this model, it is shown that the BioBERT model has a gender bias on predictions with clinical notes.

4.3.1 Limitation

Although the study was successful in validating our research question, there are still limitations that compromised the performance of the experiment. First, the sample size is still fairly small. We only obtained around 300 XML files, which produced 3,959 tokens in the training set and 2,408 in the test set. The small sample size also created a challenge when we were experimenting with different versions of the model. Even by modifying different parameters, the difference in results was negligible. If we were able to acquire a larger dataset, we expect to observe different results.

In addition, although not observed in our input, it is highly possible that one sentence contains two different pronouns. For example, 'He' prescribed medications to 'her'. Using our current labeling method, this would have been classified as both. We would need to perform additional pre-processing to exclude these sentences from the dataset, otherwise it is likely to compromise the results.

4.3.2 Future work

As NLP models are susceptible to bias, there are ways to solve this problem. To make future algorithms more accurate, multiple solutions to reduce gender bias could be implemented.

One of the methods to decrease the possibility of gender bias is to adjust the training data. The NLP algorithms are trained on data that can introduce gender bias; thus, debiasing should be one of the main approaches to improve the accuracy of models. The debiasing in terms of gender is based on gender-related word substitution. New sentences are added by altering every sentence containing gendered words, pronouns, and names, which are replaced by

opposite-gender words and entity placeholders. For instance, “Emma called her husband John” would be replaced with “NAME-1 called his wife NAME-2.” By doing this, we create a balance between both groups and make the model neutral (9).

Another method to eliminate gender bias is the gender debiasing of word embeddings. The approach consists of learning the embeddings that preserve feminine and masculine information for the words, preserve the neutrality of gender-neutral words, and eliminate gender bias from unfairly gender-biased words. After that, the decoder was used to restore the original word embeddings from the debiased embeddings (10).

One of the easiest ways to reduce the risk of gender-biased training data is by having a humans-in-the-loop. Human experts can correct algorithms errors and provide feedback that helps improve the accuracy of algorithms over time (9).

4.3.3 Consequences and Significance

Gender bias could potentially lead to different problems in algorithm application. Incorrect gender identification could misinterpret the actual proportion of males and females in the data set.

For example, gender bias could harmfully affect the statistics data used for research. Mistakes could evolve in determining conditions that are sex-specific; for instance, men tend to develop heart diseases (11), and women are likely to have major depression (12) or autoimmune diseases (13).

If the algorithm creates a gender bias, it could also make it hard to detect mistakes in assigned sex at birth made by physicians while making an entry in a patient’s note.

The algorithm could also lead to a problem capturing sexual orientation and gender identity due to the sophisticated nature of this matter. For example, the algorithm assumes that gender is binary, that gender and assigned sex at birth have a perfect correlation, and that people’s names are precise predictors of their gender or sex (14). This issue creates a space for improvements in predicting gender in the future.

5.0 Conclusion

In this study, we developed an NLP model using a pre-trained BioBERT model and performed a text classification task of gender prediction using clinical trial notes. The model’s predictions about females were only right at 35.71%. The results verified the statement in previous research that NLP models are subject to bias. This once again highlights the importance of investigating more advanced debiasing techniques.

Future work on developing NLP models for gender prediction should continue to achieve higher accuracy, which could be critical for medical research and application at the point of care.

6.0 Data and Model Sharing (Uploaded to Canvas)

The fine-tuned models are in the ‘models2’ folder. The ‘train’ and ‘test’ folders have raw data of XML files. After reading XML files and preprocessing it, the data of sentences with labels is saved in ‘train.txt’, ‘valid.txt’, and ‘test.txt’. These are saved in the same folder as the ‘BioBERT_Gender.ipynb’ file.

Reference

1. Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. Transactions of the Association for Computational Linguistics, 8, 842-866
2. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics (Oxford, England), 36(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682
3. Straw, I., & Callison-Burch, C. (2020). Artificial Intelligence in mental health and the biases of language based models. PloS one, 15(12), e0240376. https://doi.org/10.1371/journal.pone.0240376
4. Bhardwaj, R., Majumder, N., & Poria, S. (2021). Investigating gender bias in bert. Cognitive Computation, 13(4), 1008-1018
5. Bartl, M., Nissim, M., & Gatt, A. (2020). Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. arXiv preprint arXiv:2010.14534
6. 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. Advances in neural information processing systems, 29
7. Qudar, M. M. A., & Mago, V. (2020). Tweetbert: A pretrained language representation model for twitter text analysis. arXiv preprint arXiv:2010.11091
8. Stubbs, A., Filannino, M., Soysal, E., Henry, S., & Uzuner, Ö. (2019). Cohort selection for clinical trials: n2c2 2018 shared task track 1. Journal of the American Medical Informatics Association : JAMIA, 26(11), 1163–1171. https://doi.org/10.1093/jamia/ocz163
9. We can reduce gender bias in natural-language AI, but it will take a lot more work [Internet]. VentureBeat. 2020 [cited 2022 Apr 10]. Available from: https://venturebeat.com/2020/12/06/we-can-reduce-gender-bias-in-natural-language-ai-but-it-will-take-a-lot-more-work/
10. Kaneko M, Bollegala D. Gender-preserving Debiasing for Pre-trained Word Embeddings. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics [Internet]. Florence, Italy: Association for Computational Linguistics; 2019 [cited 2022 Apr 10]. p. 1641–50. Available from: https://www.aclweb.org/anthology/P19-1160
11. Bots, S. H., Peters, S., & Woodward, M. (2017). Sex differences in coronary heart disease and stroke mortality: a global assessment of the effect of ageing between 1980 and 2010. BMJ global health, 2(2), e000298. https://doi.org/10.1136/bmjgh-2017-000298
12. Albert P. R. (2015). Why is depression more prevalent in women?. Journal of psychiatry & neuroscience : JPN, 40(4), 219–221. https://doi.org/10.1503/jpn.150205
13. Angum, F., Khan, T., Kaler, J., Siddiqui, L., & Hussain, A. (2020). The Prevalence of Autoimmune Disorders in Women: A Narrative Review. Cureus, 12(5), e8094. https://doi.org/10.7759/cureus.8094
14. Asr, F. T., Mazraeh, M., Lopes, A., Gautam, V., Gonzales, J., Rao, P., & Taboada, M. (2021). The Gender Gap Tracker: Using Natural Language Processing to measure gender bias in media. PloS one, 16(1), e0245533. https://doi.org/10.1371/journal.pone.0245533
15. van Mulligen, E. M., Fourrier-Reglat, A., Gurwitz, D., Molokhia, M., Nieto, A., Trifiro, G., Kors, J. A., & Furlong, L. I. (2012). The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. Journal of biomedical informatics, 45(5), 879–884. https://doi.org/10.1016/j.jbi.2012.04.004
16. Wang, Y., Afzal, N., Fu, S., Wang, L., Shen, F., Rastegar-Mojarad, M., & Liu, H. (2020). MedSTS: a resource for clinical semantic textual similarity. Language Resources and Evaluation, 54(1), 57-72.