Differential Gender-Based Outcomes for Cardiovascular Disease: BERT & Extracting Insights on Patient Experiences from Reddit Data

Authors: Akaash Kambath, Swetha Pola
University of California, Berkeley School of Information W266 Spring 2021

Abstract

In 2021, there continues to exist disparities in diagnosis, treatment, and outcomes of cardiovascular disease for women. In this study, we develop multiple classification models and perform BERT fine-tuning for sentence classification to identify whether a post related to cardiovascular disease was posted in the subreddit r/AskWomen or r/AskMen. This classification is the first step in deep learning research on social media cardiovascular health data to extract patterns from patient-driven text, paving the way to better understand, and subsequently mitigate, gender and sex-based disparities for outcomes of cardiovascular disease.

Introduction

Background

Differential gender & sex-based outcomes for cardiovascular disease (CVD) have been brought to awareness in medical literature over the past 2 decades. Notably, a recent publication[1] conveys that on average, women will experience over 1.5x the delay of treatment, for both invasive and non-invasive treatment procedures, for cardiovascular disease. An emerging hypothesis from select domain experts is that females experience and describe CVD symptoms differently than males do. Although there has been work utilizing machine learning on structured clinical data[2], the work of applying deep learning and a text-based analysis to unstructured clinical notes, specifically for CVD, is yet to be done. Further, natural language processing (NLP) methods in the medical field have proven helpful in extracting textual fields from electronic medical records (EMR) and converting unstructured data into structured, queryable formats. However, studies utilizing deep learning methods on clinical data are sparse, due to the lack of interpretability of predictions as well as the "scarcity of publicly available [clinical] data."[3]

Social Media and Cardiovascular Health

There are several reasons to look to social media platforms to inform healthcare practices and research. Firstly, there is a cited need to go beyond EMR records in order to deepen the analysis and understanding of patient experiences of CVD symptoms and treatment. Specifically since clinical notes are both technical in nature and written by the

[1] The paper referenced here is under review for publication. The authors are UCSF researchers Dr. Maryam Panahiazar and Dr. Ramin Beygui.

[2] The study utilizes machine learning on the genome and electronic health records to predict CVD diagnosis.

[3] This paper is a literature review on NLP methods on EMR data.

secondhand perspective of a healthcare provider, deep learning performed on clinical data does not capture the nuance, language style, word choice, and other linguistic patterns of patient-driven language. Thus, we make the jump to look to social media platforms, such as Reddit, to extract patterns about patient experiences over time and outside of the doctor's office.

Secondly, not only is social media a space to make inference from, but also a space to inform and influence public sentiments, societal norms, and healthcare education. A recent paper[4] from the European Heart Journal highlights that the healthcare professionals and organizations should consider engaging in social media to revolutionize "traditional means of obtaining and disseminating medical and scientific education." Specifically, such social media efforts can "counterbalance un-reviewed and biased information" online, which the paper finds, could "reduce the burden of cardiovascular disease" in the population overall. For example, a study finds that a Facebook intervention for secondary prevention of cardiovascular disease" proves viable in older adult populations.[5] Thus, the results of assessing social media and the understanding subculture of CVD-related discussions can inform education and prevention efforts at large.

Finally, we build off of existing research, such as the study where the authors utilize real patient social media records to predict 10-year CVD binary risk scores. The classification model achieved an AUC of 0.69, suggesting a significant correlation between the linguistic features of patient social media posts and their actual cardiovascular health outcomes.

Gender and Biological Sex

Gender and sex are neither equivalent to one another, nor exist in a binary fashion. Namely, gender is an identiy and chosen; whereas sex is biological and assigned.

In healthcare settings, both gender and biological sex are important in addressing a patient's needs and populations such as those who are intersex or identify as transgender and/or gender non-conforming experience adverse healthcare outcomes at disproportionate rates. In particular, trans populations "appear to have an increased risk for myocardial infarction and death due to cardiovascular disease,"[6] which may or may not be related to gender-affirming hormone therapies.

In this project, we intentionally choose to not make an assumption about an individual's gender and/or sex, unless the gender/sex has been self-reported through the post. We frame the research question in such a way that does not force our hand in inferring or assuming an individuals' gender. This being said, we believe the results of our analysis remain to be determining steps in deepening the understanding CVD, especially when we later apply the logic of assuming a majority proportion of commenters at the group level in the r/AskWomen and r/AskMen subthreads do identify as women and men, respectively.

Research Question

We seek to answer the classification task: given a post originated from either the subreddit

[4] The authors perform a contemporary review of the role of social media in cardiovascular medicine.

[5] Study finds a viable link in social media interventions for cardiovascular disease.

[6] This review examines the relationship between hormone therapy and cardiovascular events in trans folks.

r/AskMen or subreddit r/AskWomen, can we classify which of the two it came from? Further, can we decode the neural model to interpret what the predictive textual features for each class are?

Data

Scraping Data

To product a dataset, we utilized the praw library for Python to scrape the Reddit API. Specifically, we pulled posts from the subreddits r/AskWomen and r/AskMen which included at least one of the following keywords in either the title, body, or comments of the post: 'heart surgery', 'heart attack', 'ekg', 'cardiac', 'stent', 'cholesterol', 'stroke', 'chest pain', 'bypass','sepsis','pass','coronary artery','PCI','CVD', 'cardiovascular','grafts'. This list of terms was approved as representative of XXX by medical professionals at UCSF.

Next, we cleaned the data to allow for representation from text that originated in either the title, body, or comments of the post in our final dataset, so long as it contained one one of the aforementioned terms. This was a step taken to ensure we captured the case where even if the title of the post did not reference CVD-related terms, but the comments of the post did, we removed the title text, and included the comment text. This resulted in a dataset where 1083 examples originated from r/AskWomen and 1595 examples originated from r/AskMen.

Filtering Data

As we progressed through the modeling phases, we found that the noise in some of the example data was too extraneous to include in our dataset. We utilized regex along with manually scanning and filtering out noisy examples. This resulted in a final dataset where 697 examples originated from r/AskWomen and 748 examples originated from r/AskMen.

Labeling

To label the dataset, we labeled all posts originating from r/AskWomen as examples of the positive class and all posts originating from r/AskMen as examples of the negative class.

We choose this approach for the sole reason to avoid making inference about the gender of the author of a post. Similar works[7] have made assumptions that the top replier of r/AskWomen is usually a woman and analogously for r/AskMen to label the gender associated with a post, though to avoid entangling ourselves with such an assumption, we decided to attempt to solve the problem of classifying which of the two subreddits a post originated from. In our findings, we recommend that analyses of the genders associated with each of the subreddits be done with caution.

Modeling Approaches

The metrics used to evaluate each model were precision, recall, F1 score, and AUC.

Logistic Regression

In our initial modeling phase, we experimented with utilizing a Logistic Regression, in order to supplement interpreting results of the neural network model. We trained the logistic regression model on n-grams to predict the class of the examples. An input of around 30,000 bigrams, trigrams, and

[7] This article goes through an attempt to predict gender from Reddit posts.

4-grams resulted in a model that was unable to converge; fine-tuning hyperparameters did not help.

We then layered in a feature that captured the sentiment of the post. Surprisingly, trained on just this one feature, the model performed at 100% accuracy on both the train and validation sets.

Ultimately, although we are not able to use the model to cross validate the results from Lime, we learn that sentiment is a key predictor of the class, and that the solution space is too complex to model with a regression.

Softmax Classifier with Word2Vec

As a simple approach to deep learning, we developed a 2 layer neural network, with one hidden layer and a softmax layer. The network was trained on context-free Word2Vec embeddings. This model serves as the baseline.

CNN with Word2Vec

Next, we developed a convolutional neural network, with one convolutional layer, trained on context-free, 300-dimensional Word2Vec embeddings. This model, despite cross-validating hyperparameters, led to excessive overfitting and poor performance.

BERT with Softmax

Using 768-dimensional pre-trained embeddings from BERT, we performed fine-tuning to apply BERT to our sentence classification task at hand. We referenced the example notebook on TF hub for BERT with sentence classification[8] to complete this task.

[8] Predicting movie sentiment with BERT (github repo)

Furthermore, we performed robust data filtering to manually remove noisy posts, which we found were flagged through our keywords search, but unrelated to CVD. This improved the clarity of the trained embeddings, and the performance of the model as a whole.

Results

Table 1: Precision, Recall, F1

Model	Precision	Recall	F1-score	AUC
Logistic Regression*	1.00	1.00	1.00	1.0
Softmax (context-free)	0.56	0.56	0.56	0.60
CNN (context-free)	0.58	0.56	0.53	0.57
Softmax (BERT)	0.73	0.69	0.71	0.72

*Note: the logistic regression results are reported for the model in which only sentiment is a feature.

Interpreting Results with LIME

Perhaps the most crucial component of this study is utilizing LIME to demystify and interpret the weights of the neural network. In this section we explore which features were most predictive of each class. We also look at examples labeled incorrectly by the classifier to understand which word features are perhaps, overlapping in nature between the 2 classes. Although a few samples of examples do not convey the full innerworks of learned parameters, they do help provide insight. For example, this deeper look at the false negatives and false positives may demonstrate that the model has learned the difference between men's and women's symptoms.

True Negatives

In Figures 1a-1z are examples of texts which were true negative (labeled r/askMen, predicted r/askMen). The tokens that came back as most predictive for this post are: "id," "heart," "ended." Note that we removed the stopword "i", so "id", or "i had" is the next best token that denotes the post is written from a 1st person point of view.

False Negatives

In Figures 2a is an example of text which is a false negative (labeled r/askWomen, predicted r/askMen). Figure 2a demonstrates a case where although the post originated in r/askMen, the content of the post is about the author's father's heart attack. This may or may not demonstrate the model picking up on men's symptoms.

True Positives

In Figure 3a is an example of text which is a true positive (labeled r/askWomen, predicted r/askWomen). Figure 3a shows the authors' experience of heart attack post-pregnancy; Lime shows that "pregnancy" was the strongest predictor of the positive class. It's also interesting to note the word "felt" is used repeatedly in this post, and contributes to its classification as well.

False Positives

In Figures 4 are examples of text which is a false positive (labeled r/askMen, predicted r/askWomen). Figure 4a demonstrates a case where although the post originated in r/askMen, the content of the post is about the author's mother's heart attack. This may or may not demonstrate the model picking up on womens' symptoms.

Discussion

In this study, we find salient textual features which are predictive of a post having originated in r/AskWomen and r/AskMen. We extrapolate that some of the features in r/AskWomen, as discussed previously, are associated with symptoms and experiences closely aligned with women and cardiovascular disease (and vice versa for r/AskMen). Although not all authors of posts may be women or men, by framing the results as class features, we begin to get at the gendered and sex-based differences of CVD, insofar as the data, model, and asusmptions allow us to do. As far as future work goes, we see a number of directions for next steps.

Firstly, due to default limits set by the Reddit API, we were unable to scrape beyond the most recent 1000 posts for each subreddit that satisfied the terms we were searching on. A future endeavor may find a way to work with Reddit to allow us to scrape beyond the most recent 1000 posts. This may prove to be useful because when manually reviewing the 2000 posts, there was a significant amount of noise in the data that had to be removed from the dataset. Thus by nullifying the limit, we can ensure we have a substantial amount of robust data to work with after the noise has been filtered.

Another technical endeavor worth looking into in the future may be to attempt to fine-tune BERT models that are pre-trained on clinical notes or biomedical papers, such as BioBERT and ClinicalBERT. This may provide better results on Reddit posts containing a decent amount of medical terms.

Secondly, we suggest a longitudinal context analysis study to look at Reddit users' posts over time, as this would provide a glimpse of how CVD symptoms may manifest or change over time for individuals of different genders.

Thirdly, we reflect on how users of varying demographics may turn to online resources for medical help at greater rates than other demographics, perhaps due factors such as healthcare access barriers. Extending the work to contextualize for race, class, ability, and so on, can also provide deeply valuable insights about the environmental, medical, biological, and structural factors that make for varying outcomes of cardiovascular disease.

Overall, we believe that through deeper extraction of features from social media platforms, beyond Reddit, we can provide substantial evidence for change in healthcare guidelines. One potential outcome of future research may be: if a non-dominant or gender/sex-specific symptom for CVD (as informed by social media analysis) is flagged during a clinical visit, to follow up with that patient on expedited cardiovascular disease diagnosis & potential treatment.

Appendix

Figure 1a.

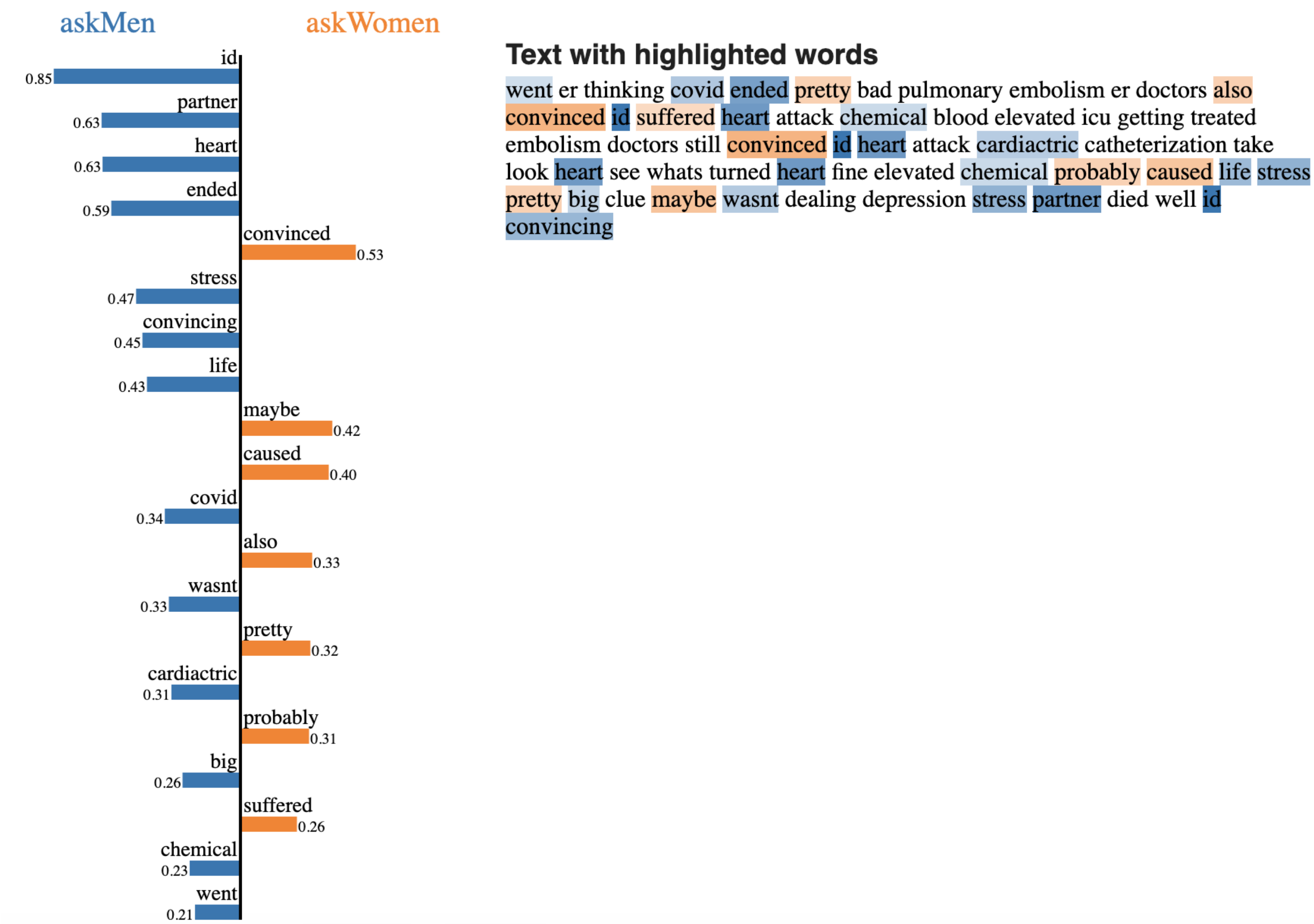

Figure 2a.

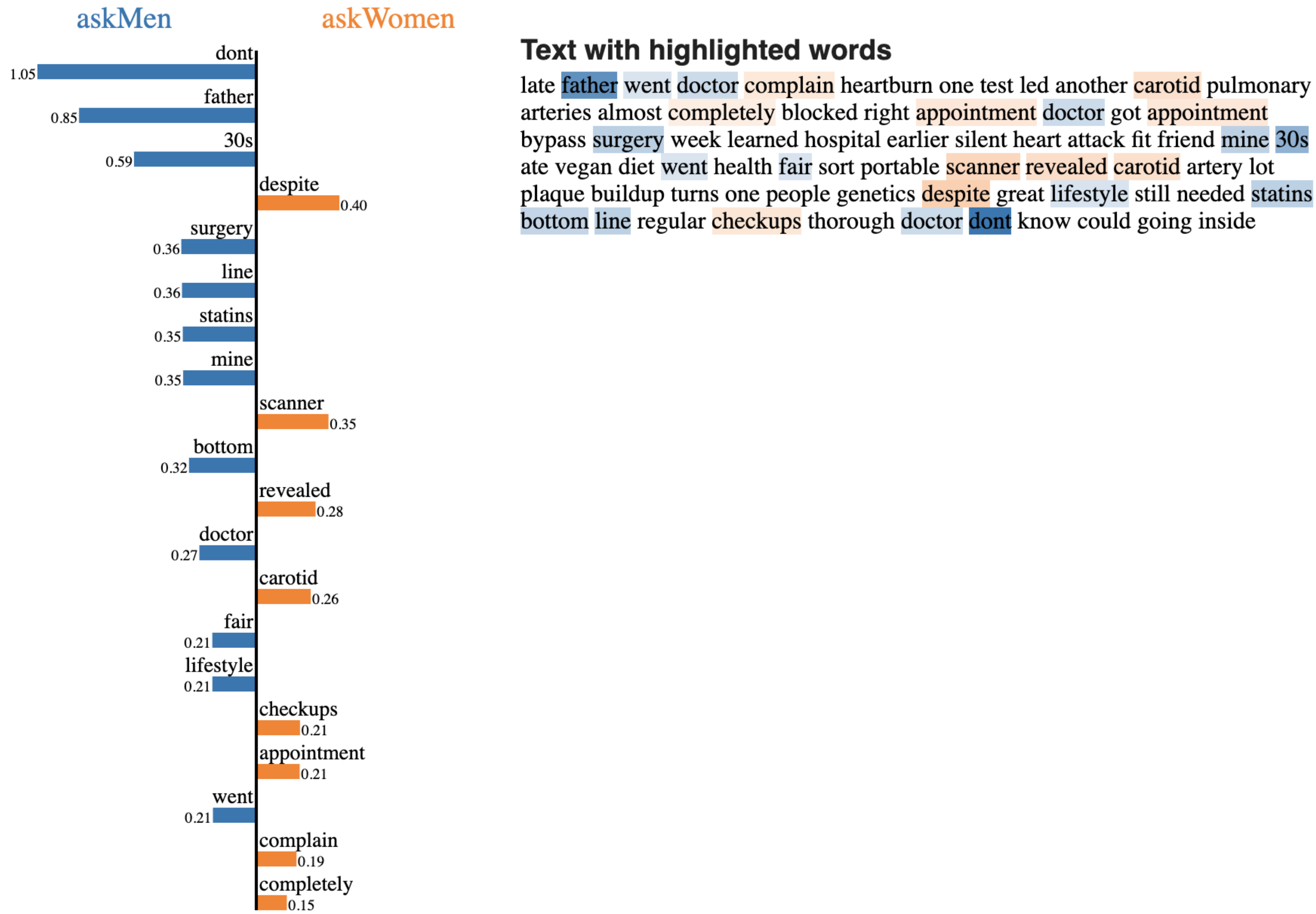
askMen
askWomen
dont 1.05
father 0.85
30s 0.59
despite 0.40
surgery 0.36
line 0.36
statins 0.35
mine 0.35
scanner 0.35
bottom 0.32
revealed 0.28
doctor 0.27
carotid 0.26
fair 0.21
lifestyle 0.21
checkups 0.21
appointment 0.21
went 0.21
complain 0.19
completely 0.15
Text with highlighted words
late father went doctor complain heartburn one test led another carotid pulmonary arteries almost completely blocked right appointment doctor got appointment bypass surgery week learned hospital earlier silent heart attack fit friend mine 30s ate vegan diet went health fair sort portable scanner revealed carotid artery lot plaque buildup turns one people genetics despite great lifestyle still needed statins bottom line regular checkups thorough doctor dont know could going inside

Figure 3a.

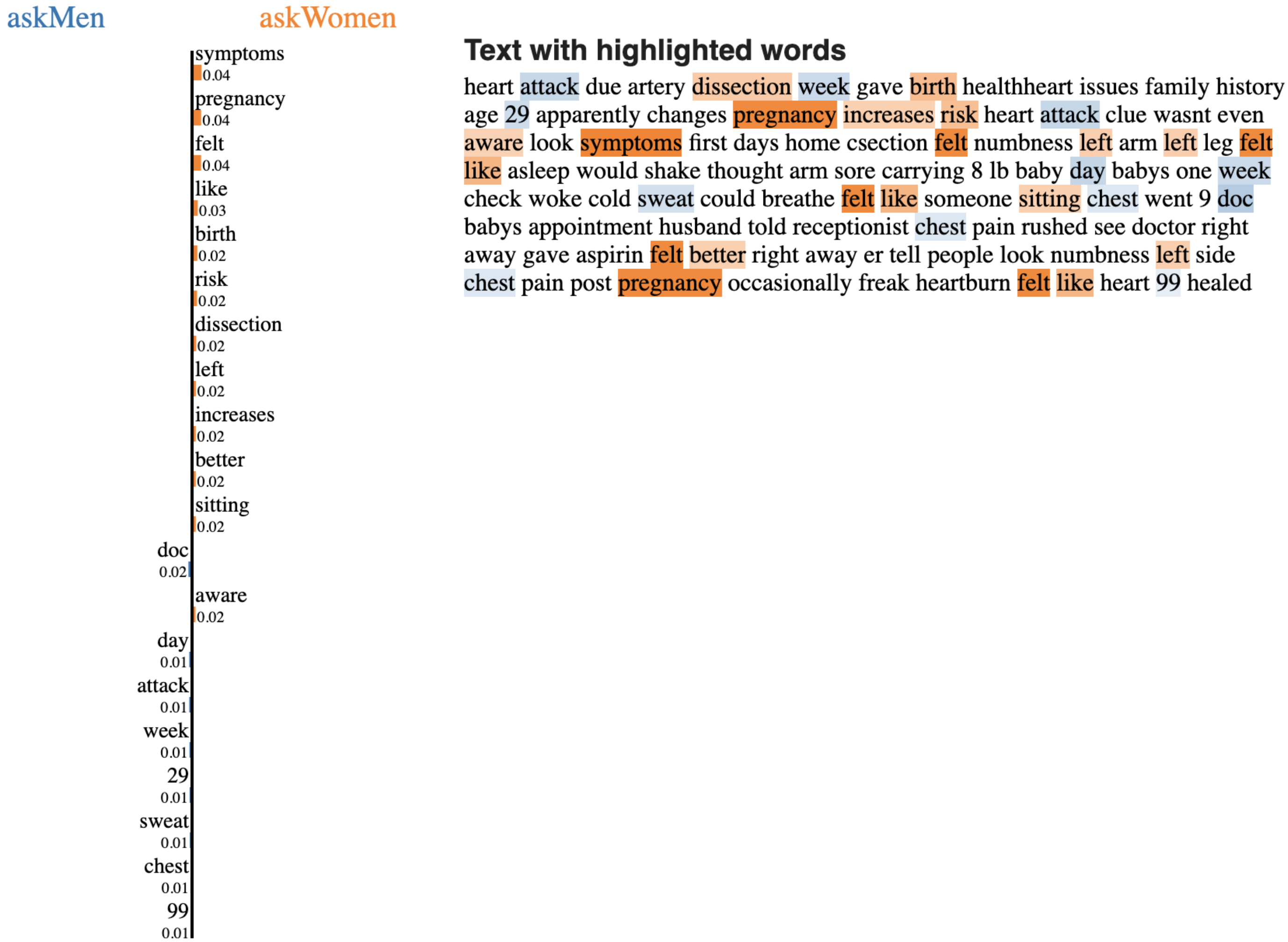
askMen
askWomen
symptoms
0.04
pregnancy
0.04
felt
0.04
like
0.03
birth
0.02
risk
0.02
dissection
0.02
left
0.02
increases
0.02
better
0.02
sitting
0.02
doc
0.02
aware
0.02
day
0.01
attack
0.01
week
0.01
29
0.01
sweat
0.01
chest
0.01
99
0.01
Text with highlighted words
heart attack due artery dissection week gave birth healthheart issues family history age 29 apparently changes pregnancy increases risk heart attack clue wasnt even aware look symptoms first days home csection felt numbness left arm left leg felt like asleep would shake thought arm sore carrying 8 lb baby day babys one week check woke cold sweat could breathe felt like someone sitting chest went 9 doc babys appointment husband told receptionist chest pain rushed see doctor right away gave aspirin felt better right away er tell people look numbness left side chest pain post pregnancy occasionally freak heartburn felt like heart 99 healed

Figure 4a.


askMen
askWomen
mum 0.84
hey 0.39
symptoms 0.37
heart 0.26
mother 0.24
wasnt 0.23
strong 0.22
hard 0.19
quick 0.18
note 0.17
didnt 0.16
support 0.15
shes 0.15
condition 0.14
signs 0.14
pain 0.13
stress 0.13
aware 0.12
super 0.12
drive 0.11
Text with highlighted words
hey quick note super capable strong hard working mother heart attack months ago please tell mum aware signs way get help mum wasnt case clogged arteries stress triggered underlying heart condition didnt know collapse left arm hurting instead terrible indigestion short breath terrible chest pain could barely walk flight stairs shes nurse able recognize sister drive hospital despite typical symptoms please keep sweet momma safe support best

References

[1] This study is under review; the authors are UCSF researchers Dr. Maryam Panahiazar and Dr. Ramin Beygui.

[2] Li, Jingjing, et al. "Decoding the Genomics of Abdominal Aortic Aneurysm." *Cell*, vol. 174, no. 6, 21 July 2018, doi:10.1016/j.cell.2018.07.021.

[3] Sheikhalishahi, Seyedmostafa, et al. "Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review." *JMIR Medical Informatics*, vol. 7, no. 2, 2019, doi:10.2196/12239.

[4] Ricardo Ladeiras-Lopes, Lavinia Baciu, Julia Grapsa, Afzal Sohaib, Rafael Vidal-Perez, Allan Bohm, Harri Silvola, Maria Rubini Gimenez, Saverio Muscoli, Markus Wallner, Amina Rakisheva, Vivien Klaudia Nagy, Martin R Cowie, Sarah C Clarke, Stephan Achenbach, the 'Cardiologists of Tomorrow', Digital Health and Media Committees of the European Society of Cardiology, Social media in cardiovascular medicine: a contemporary review, European Heart Journal - Digital Health, Volume 1, Issue 1, November 2020, Pages 10–19, https://doi.org/10.1093/ehjdh/ztaa004

[5] Partridge, Stephanie R et al. "Cardiac Patients' Experiences and Perceptions of Social Media: Mixed-Methods Study." Journal of medical Internet research vol. 19,9 e323. 15 Sep. 2017, doi:10.2196/jmir.8081

[6] Irwig, Michael S. "Cardiovascular health in transgender people." Reviews in endocrine & metabolic disorders vol. 19,3 (2018): 243-251. doi:10.1007/s11154-018-9454-3

[7] Fleming, Eamon. "Predicting Gender in Reddit Comments." www.eamonfleming.com/projects/reddit-gender.html.

[8]https://github.com/google-research/bert/blob/master/predicting_movie_reviews_with_bert_on_tf_hub.ipynb

[9] Mosca, Lori et al. "Sex/gender differences in cardiovascular disease prevention: what a difference a decade makes." *Circulation* vol. 124,19 (2011): 2145-54. doi:10.1161/CIRCULATIONAHA.110.968792

[10] Shah, A.D., Bailey, E., Williams, T. et al. Natural language processing for disease phenotyping in UK primary care records for research: a pilot study in myocardial infarction and death. J Biomed Semant 10, 20 (2019). https://doi.org/10.1186/s13326-019-0214-4

[11] Trivedi, Hari M., et al. "Large Scale Semi-Automated Labeling of Routine Free-Text Clinical Records for Deep Learning." *Journal of Digital Imaging*, vol. 32, no. 1, 2018, pp. 30–37., doi:10.1007/s10278-018-0105-8.

[12] https://github.com/gaphex/bert_experimental