Gauge against the machine: improving representations within sociotechnical instruments to enrich context and identify biases

A Dissertation Presented

by

Joshua R. Minot

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
Specializing in Complex Systems and Data Science

May, 2022

Defense Date: June 1st, 2022
Dissertation Examination Committee:

Peter Sheridan Dodds, Ph.D., Advisor
Christopher M. Danforth, Ph.D., Advisor
Donna Rizzo, Ph.D., Chairperson
Randall Harp, Ph.D.
Jeremiah Onaolapo, Ph.D.
Cynthia J. Forehand, Ph.D., Dean of Graduate College

Abstract

The proliferation of digital data across all areas of society has transformed our ability to hypothesize, study, and understand social systems. From this richness of data we have seen the development of innovative instruments to study—and make decisions with—the digital artifacts of the modern day. These developments build on advancements in computation, connectivity, analytical methodologies, and sociological theories. The sociotechnical instruments we have developed have been revolutionary to how we understand society and how we conduct business, but with these broad leaps comes ample room (and need) for more nuanced advancements. As with the development of any field, as the digital humanities evolve there is opportunity for targeted progress and the need for more tectonic shifts in practices. Iterative improvements include building more full-featured instruments that include a broader set of variables when analyzing and presenting results. More profound topics such as fairness, accountability, transparency, and ethics need increased attention as well—especially to create equitable, pro-social tools. Both in academia and in industry, there is room to improve how we curate, study, and operationalize data sets and the AI pipelines that sit atop them. Here we use natural language processing, machine learning, tools from data ethics, and other methods to explore how we can contextualize results and improve representations within instruments used to understand sociotechnical systems.

In the first study we examine the dynamics of responses to posts by US presidents on Twitter. These results offer a piece of culturally significant data in themselves—the ratio of response types is an unofficial measurement on the platform. Moreover, the results improve our understanding of the temporal dynamics that lead to the final counts that users may ultimately see. Deeply analyzing response activity dynamics provides insights on how the public responds to posts, the tenacity of supporters, and abnormalities that may be indicative of inauthentic behavior.

The second study examines the interaction between gender biases in health records and language models and how to mitigate these biases. We present specific language that is more commonly associated with female and male patients. We go on to demonstrate how the deliberate augmentation of text can minimize the gender signal present in data while retaining performance on medically relevant tasks. We conclude by showing how much of this bias is domain specific, and the non-trivial interaction with general-purpose language models.

Our final study investigates gender bias in resume text and relates this bias to the gender wage-gap. We show that language differences within occupations are associated with the gender pay gap. Our results highlight the value of utilizing high dimensional representations of individuals and the potential for previously undocumented biases to influence hiring pipelines.

Material from this dissertation has been published in the following form:

Minot, J. R., Arnold, M. V., Alshaabi, T., Danforth, C. M., Dodds, P. S. (2021). Ratioing the President: An exploration of public engagement with Obama and Trump on Twitter. PloS one.

Minot, J. R., Cheney, N., Maier, M., Elbers, D. C., Danforth, C. M., Dodds, P. S. (2022). Interpretable bias mitigation for textual data: Reducing genderization in patient notes while maintaining classification performance. ACM Transactions on Computing for Healthcare.

Minot, J. R., Maier, M., Cheney, N., Demarest, B., Danforth, C. M., Dodds, P. S., Frank, M. R. (2022). He said, she read: Job-specific language is associated with gender pay gap, while semantic biases are associated with employment gap. *Manuscript currently under preparation.*

Acknowledgements

None of this could have possible without the support those around me. When I entered UVM I thought I knew the course my graduate studies would take—and I am glad that my experience turned out differently. I never could have imagined how the driven, quirky, inquisitive, and caring people I would come to know could shape my trajectory in such an amazing way. As with any experience, there are peaks and valleys (but far more peaks in this case). I am grateful to all who have been a part of this experience. I am where I am today in part because of those who I have had the privilege of sharing this time with. So with that, I have a few people to thank.

First, Peter and Chris for introducing me to this area of work and building an incredible program that I could never leave after just 2 years. My experience, and that of many others, has been greatly benefited by your dedication to leadership through kindness, curiosity, and passion.

Melissa Rubinchuk has always been there support me in mind, body, and spirit.

Donna Rizzo for welcoming me to UVM and providing me with my first research opportunity (and early guidance on navigating life as a wee graduate student).

Randall Harp has enriched my experience at UVM with ever-fascinating discussions that have shaped my views of ethics—both in and out of the data science realm.

Jeremiah Onaolapo has always entertained expansive and thought provoking conversations that helped me think through the messiness of our modern world.

Juniper Lovato and Laurent Hébert-Dufresne provided advice and fostered the kind of community I am proud to be a part of.

My undergraduate advisors, Mark Feinstein and Larry Winship, entertained wild research projects and planted the idea of graduate studies.

John Ring taught me enough to be dangerous on the VACC (and the joys of picking up heavy things).

Mariah Boudreau, Erin McBride, Bryn Loftness, Erik Weis, Milo Trujilo, Mitchell Joseph, Han Yin Cheng, and Irfan Tahir are the best support crew one could hope for as we emerge from many months of isolation and rebuild our worlds.

Quin Mann provided relentless encouragement in the final miles.

Micheal Arnold entertained wild research ideas (and early stage, hands-on hardware adventures).

Peter Larsen powered through late nights at the whiteboard during my first year at UVM.

Morgan Frank has fostered new research opportunities and always had helpful advice for navigating graduate school while staying human.

Marc Maier has brought new perspectives and patient feedback.

John Meluso for thoughtful research conversations on life in academia (and beyond).

Sophia Hodson for thinking about how we can make the world a little better.

Nicholas Cheney, Danne Elbers, Bradford Demarest, Andy Reagan, Ryan Gallagher, Jesse Gourevitch, and Robert Worley for excellent collaborations and conversations.

To Jane Adams, Kesley Linnell, Mikaela Fudolig, Amelia Tarren, Colin Van Oort, Kelly Gothard, David Dewhurst, Thayer Alshaabi, Max Green, Anne-Marie Stupinksi, and Vanessa Mayhaver for a vibrant community.

Finally, thanks to my parents whose unwavering cheerleading and enduring support helped me embark and stay on this path.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

In the 21st century, most fields have experienced an explosion of available data. Extracting meaning from vast volumes of digital artifacts has become a task relevant to research, business, and policy. Sources include trace data—such as social media and cell phone activity—that provide both a conduit for communication and an instrument for sensing individual behaviors. Other more deliberate collections of human activities and attributes include electronic health records, worker resumes, and libraries of digitized books—artifacts that are constructed with a clear purpose, but never-the-less contain nuanced and often times difficult to extract insights. Further adding to the complications, many pieces of information reside in unstructured text data. With the advent of large-language models, the potential of text data is greater than ever. Impressive advancements have been made but limitations are becoming more apparent. There is the need for progress in under-addressed areas if we want to see the ethical and sustainable maturation of how we utilize sociotechnical data.

Two areas of focus include 1) improving our understanding of biases in these datasets, models, and analytical pipelines; and 2) finding new ways to provide context to large data sets, enriching our understanding of human behavior as sensed through the digital lens.

Improving representations is a key component of text-as-data research. Often times these datasets—such as social media, medical records, and resumes—are viewed in aggregate without interrogating contextualizing information. High-level views of these systems are a vital starting point and important perspective to have—and require non-trivial research and engineering efforts. Occam's razor is indeed a worthy guiding principle. We should not needlessly complicate our analyses. We need to take system level readings and understand the macro-scale to properly understand the meso- and micro-scale phenomena. However, there are justified cases for additional complexity. Some of the most pressing reasons for exploring more nuanced representations come from the ethical considerations present in society increasingly defined by artificial intelligence (AI) and big data. Questions such as, who gets to decide when systems, groups, or individuals are adequately represented? What questions do we want to ask? What level of performance is sufficient and for what groups?

Accurate representations of individuals and groups is a broad issue in computational social science and machine learning. From a technical perspective, the task may fall under the umbrella of feature engineering and data set curation—crucial tasks familiar to most practitioners of AI and data science. From a philosophical perspective, the question becomes more abstract. What constitutes an individual? What information must be included to accurately represent a given person, case, or state in an accurate fashion? To answer these questions we engage with the perspec-

tive broadly offered by the algorithmic and data fairness space. Fairness in AI is itself asking seemingly basic and foundational questions with profound implications. There are lofty issues raised (such as asking society to neatly define a set of ethical objectives) while more earthly questions abound as well (e.g., does the algorithm do what we think it does?).

In the 21st century digital trace data abounds and we are inclined to 'look where the light is'. Platforms that were not designed as research instruments are being repurposed to yield insights about society that were previously unimaginable. At the same time, many of these platforms *were* designed to market to the masses—influencing brand opinions of politicians, products, and policies. In light of the proliferation of these approaches, now more than ever is the time for more critical examinations of these data and analytical frameworks [4]. This includes creative approaches to extracting contextualizing information from the platforms that generate this data (e.g., Chpt. 2), understanding how large language models interact with societal bias (e.g., Chpt. 3), and investigating the impact of latent factors in high dimensional data (e.g., Chpt. 4).

Throughout this piece, textual data will play a central role. Text-as-data is a common analytical paradigm for investigating sociotechnical systems. Repositories of textual data are used to understand the specific system under examination, or as a means to an end when capturing latent attributes of natural language (e.g., training language models). Language data is inherently high dimensional and contains some of the richness that is lacking in other data streams—with this comes the potential for biases to be incorporated in opaque manners. Further, the sheer volume of data used in many cases makes careful curation of text data difficult.

The big data revolution has been eating many unsuspecting fields. Articles such as *The unreasonable effectiveness of data* (2009) [5] give credence to the now deeply held view that more data is better in most fields. Indeed, the presence of large data sets and the infrastructure to develop large models has revolutionized at least corners, if not the entirety, of many fields. The hyper-scale data set revolution has perhaps had the most striking impact in the field of natural language processing. By the mid-2010s the word-embedding revolution had firmly established its worth—with models such as word2vec [6] and GloVe [7] demonstrating an impressive ability recover semantic information from what in retrospect is a relatively straightforward model architecture and modest training data set. Massive (but minimally curated) data sets such as the Colossal Clean Crawled Corpus [8] (750GB of "clean" English text) has enabled the creation of large models with billions of parameters such as T5 [9], GPT-3 [10], and OPT [11].

It is important to think critically about the generative source of the data used for directly analyzing sociotechnical systems and training these models. For instance, in the case of social media, the fact that a communication is written by bots, anti-social actors, partisans, etc. may endue text with dimensions that are not readily ascertained by the text of the artifact itself.

There are a plethora of factors that lead to sub-optimal representations in data sets, some of which can never be fully reduced. As Dana Coyle writes in *Socializing Data* (2022) [12],

> Although big data offers great potential for progress, any data set is a limited, encoded representation of reality, embedding biases and assumptions, and ignoring information that cannot be codified.

There are limiting factors that we have yet to look at and there are blindspots that may be irreducible. This dissertation in part demonstrates some of the views we can take to improve our understanding of data sets that describe sociotechnical phenomena and the instruments we build atop these data sets. We show how we can further enrich social media posts with information from the broader user base—lending a voice to how the broader platform responds to famous individuals. We investigate how demographic signals are embedded in textual data, how these signals interact with large language models, and what can be done to mitigate bias detected through this lens. Finally, we present a case where we extract salient information from the text data of resumes, highlighting the signal present in this data and how it can help us understand real world harms. Taken together these approaches highlight case studies where researchers and engineers alike can dive deeper and seek to understand salient features of data and analytical pipelines.

1.2 Background

Natural language processing

Natural language processing (NLP) is concerned with deriving representations of language that can be utilized in computational pipelines. The field sits at the intersection of computer science and linguistics, and has experienced catalyzing advancements in recent years—owing to synergies in the development of larger-datasets, greater computational power, and advanced language models. As most relevant to this piece, we focus on topics such as token and concept representation, distributional measures, text classification, and language models.

Text representation

Free text is inherently a high dimensional artifact. Even without factoring in contextualizing information and metadata, language and the semantic information it contains is highly complex. Representing text in a fashion that enables statistical analysis and the application of AI requires careful consideration.

Starting with tokens—the atomic unit of language in an NLP pipeline—we can then build up broader representations of documents. The n-gram—often a space separated sequence of characters—is a conceptual building block for most (but not all) parsers and a small unit of language that humans can reason about. The n-gram is also a common token for simple language models. It is worth noting that n-grams may refer to space separated character sequences (e.g., one word is a 1-gram), sub-word sequences, or individual characters, depending on the context. For the purpose of this dissertation, when referring to n-grams we are discussing space separated character sequences unless otherwise stated. Defining the n-gram can be a nuanced task with *scriptio continua* languages which do not incorporate spaces (such as Chinese, Japanese, and Thai). More generally, in the era of widely used unicode characters—especially emoji—discretion is required when deciding what constitutes an n-gram. Often, varying N is helpful when analyzing a given corpus—in general higher order n-grams might reveal meaningful sequences such as proper nouns and phrases.

It is generally at the n-gram level that one might implement normalization, stemming, and/or lemmatization. All of these efforts are motivated by mapping n-grams that represent the same concepts to the same character representations. For a basic example, we might map the plural "computers" to the singular "computer". These

efforts are well motivated, but for the research outlined in this dissertation we rarely employ normalization efforts. There are a few main reasons for avoiding normalization in our research. For one, the volume of data is often sufficient that the main signal associated with a given concept is readily detected even if there are multiple tokens associated with that concept. Secondly, and especially in social media analysis, keeping the original text intact can be informative for downstream analyses—both because misspellings, specific tenses, and similar can be meaningful, but also because it can be difficult to fully anticipate the breadth of possibilities with technical jargon and 'internet speak' (which can have non-trivial interactions with normalization efforts).

Tokenizing supports many downstream tasks, such as direct analysis of language distributions and training language models. For the former task, tokens are often kept relatively whole (e.g., complete n-grams). For the later task of language modelling, some of the earlier word-embedding models (e.g., GloVe [7] and word2vec [13]) use whole 1-grams. More recent neural-network based models use tokenizers that both retain some full 1-grams and split other words into pieces (sub-strings). For instance, the Bidirectional Encoder Representations from Transformers (BERT) language model uses the WordPiece tokenizer. Other tokenizers operate on bytes as their primary unit of string representation in order to limit vocabulary sizes (e.g., Byte-level Byte-Pair Encoding [14]). Operating closer to the character level can reduce the dependency of the resulting tokenizers characteristics on the training data. Sentence level encodings have also been developed that address issues with *scriptio continua* languages. All of these tokenizers must be trained to optimize the vocabulary and have some reliance on the base training set. This is a potential source of bias, for in-

stance uncommon names are less likely to be tokenized as singular word pieces which can lead to these uncommon words be grouped together in some tokenizers [15].

While we do not discuss the entire language modelling pipeline in technical detail, we make a point to provide an overview of tokenization because the step sits at the intersection of human- and machine-interpretable data while also providing an example of how modelling assumptions can begin to instill biases in a given pipeline.

The simple power of language distributions

Taking a bag-of-words approach to a corpus and tabulating simple counts of n-grams allow for the construction of language distributions which can unlock powerful and interpretable insights. In the bag-of-words paradigm, corpora are represented by unordered sets of their constituent n-grams and their respective counts. The resulting distribution is a Zipf distribution for natural language [16]—a special case of the power law distribution where a word's frequency is inversely proportional to its rank. The Zipf distribution parameters in themselves can be informative for estimating coarse-grained insights about a corpus—for instance deviations in the α scaling parameter may be indicative of errors in data processing or the presence of unnatural language. But much of the power in tabulating corpora as Zipf distributions comes from investigating the constituent elements. The ranks and normalized frequencies of n-grams can provide contextualized information on the relative prevalence of terms and phrases. Arranging collections of these distributions in the temporal domain creates time series that provide insights on the dynamics of the foundational building blocks of narratives in a given system. Alternatively, we can compare distributions to quantify differences in the overall distribution and specific elements that have the highest

contrast in the two systems. Using information theoretic (and similar) approaches—such as Kullback-Leibler divergence [17], Jensen-Shannon divergence [18], and the recently introduced Rank-turbulence divergence [2]—we can calculate the degree to which language distributions diverge. From this calculation, we can identify the individual n-grams contributing to this divergence—a powerful tool for characterizing each distribution relative to each other.

Language models

Language modelling can be thought of in a very broad sense as the process of creating representations of language that machines can meaningfully reason about. That is to say, the representations of the text retain semantic meaning so that useful tasks can be carried out in computationally efficient manners.

Recent advancements in language models have revolutionized the way we view text data. Larger models, more extensive pre-training, domain-adaptation, and more user-friendly libraries have lead to a proliferation of language models for a multitude of tasks. In the present dissertation, we utilize language models for classification, establishing semantic similarity, and evaluating bias.

Language models generate word-embeddings, a term that we will use to discuss the specific, numeric output of language models. Word-embeddings allow us to calculate meaningful results from numeric representations of words (e.g., cosine similarity) and were in the past perhaps more central to the use of language models. It is worth noting that new paradigms with large-language models and user friendly model application program interfaces (APIs) have partially abstracted the role of the embedding away from the NLP practitioner (for instance, zero-shot learning may largely be concerned

with text-to-text interfaces). Of course, regardless of downstream task language models produce some form of numeric representation—what varies is the extent to which the produced embedding is optimized for certain tasks, such as evaluation of semantic similarity.

Language models generate word embeddings by learning the distribution of language for a given training data set. At a high-level, most language models are trained—at least in part—by predicting words or tokens based on some context language. Stated more generally, language models learn what language commonly occurs together. Classically, the distributional hypothesis states "a word is characterized by the company it keeps" (Firth, 1957) [19]. The "company" or context can be constructed in a few different ways. When representing documents or context, one of the simplest approaches is the bag-of-words (BOW), a representation of language that ignores the order and proximity of tokens. A BOW representation is one of the ways the now classic word2vec language model can be trained [6]. Another technique developed for word2vec is the skipgram—a training procedure that considers relative position of context and target words, down weighting the context terms that are further away. Both techniques produce word2vec word embeddings that are *static* in that after training, when retrieving the embedding for a word it will not change based on the context (i.e., embeddings are retrieved much like a lookup table). There are variations on the skipgram approach, such as the fastText model [20] that includes character-level n-grams to improve representations.

A major advancement in language modelling came with sequence-to-sequence models in 2014 [21]. In brief, sequence-to-sequence models take into consideration the previous portions of a passage—beyond the simple rolling window or BOW ap-

proach used in word2vec. One of the most significant advances with sequence-to-sequence models came with attention in 2014 [22] and the subsequent transformer model architecture of 2017 [23]. At a high level, the transformer architecture enables models to incorporate non-local interactions into the language modelling process—providing a mechanism to focus model attention on salient interactions in text. Examples of these models include Bidirectional Encoder Representations from Transformers (BERT) [24] and Generative Pre-trained Transformers (GPT-n) [25]. Owing to their use of broader sequence information, these models result in *contextual* word-embeddings where the numeric representation of a given word or token is dependent on the broader context within which it appears. Contextual word-embeddings achieve state-of-the-art performance on common benchmarks, but present challenges with interpretability and direct comparison with static embeddings [26]. The state-of-the-art models that produce these contextual word-beddings can have over 100-billion parameters [10] making the models' behavior difficult to understand.

For all of these language models (and machine learning models in general) the training data matters immensely for the performance characteristics of the model. The curation of massive data sets such as the colossal clean crawled corpus [9]—a data set created by scrapping much of the known internet and weighing in at over 25 terabytes in its fullest version—has been central in creating models with hundreds of millions or hundreds of billions of parameters. As we will discuss below, how these data sets are created and what data is included in them can be problematic from a fairness and ethical standpoint. In light of the tendency for language models to be heavily influenced by the characteristics of their training data, techniques such as transfer learning and domain adaptation have become helpful in building models

tailored to specific applications without having to retrain from scratch. Transfer learning is the use of model weights from one training context for use in another, basically initializing the model with previously learned parameters before continuing with training or deployment. In the case of BERT, the model has been adapted using transfer learning to improve performance on scientific literature [27], medical notes [28], and COVID-19 content on Twitter [29].

The above overview of some key aspects of modern language modelling is meant to highlight a few areas for potential bias to enter the process. Language models are increasingly powerful and widely used but also growing in their complexity and lack of transparency. Starting with the training, there are potential issues with the distribution of language included (e.g., hate speech, inaccurate information, personally identifiable information). This training data is then used to create tokenizers that may begin biasing language models (although, this has been addressed somewhat by more modern tokenizers). The training distribution is then instilled in the model itself, where larger and larger models make auditing and bias detection increasingly difficult. Language models are a central tool in the modern digital ecosystem—they unlock an immense amount of information contained in unstructured language—but we must seek to understand them if we wish to use them in equitable and ethical fashions.

1.2.1 Algorithmic fairness

Work in the AI fairness, accountability, transparency, and ethics (FATE) space goes by many names and has come to include researchers from a variety of disciplines. The field has tied together stakeholders from the social sciences, computer science,

philosophy, law, and others in an effort to understand how AI interacts with social systems and ideally how to mitigate harms. FATE raises questions related to societal morals and goals while forcing us to operationalize specific metrics to describe these ideals. Fairness includes critically evaluating the constructs we create to model the world and track outcomes [30]. For instance, what is the purpose of the criminal justice system, and how do we measure the stated purpose? If we say we want equitable employment outcomes, for what groups or individuals are we specifically addressing? How will equity be defined? From questions like these rise not just technical challenges but ethical and political debates.

In many regards, FATE is just good science. FATE encourages AI researchers and practitioners to think about foundational questions that may get overlooked when trying to simplify problems and how we frame them. In 2022, the National Institutes for Standards and Technology (NIST) released a publication entitled *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence* [31]. The publication states "Trustworthy and Responsible AI is not just about whether a given AI system is biased, fair or ethical, but whether it does what is claimed." The authors go on to state "What is missing from current remedies is guidance from a broader SOCIO-TECHNICAL [sic] perspective that connects these practices to societal values." Indeed, to address FATE issues in the fullest we need a full sociotechnical accounting of the system under examination. There is a conversation between the technical and social side of FATE. The social perspectives push inquiries into the performance of AI—inquires that must be informed by real world impacts. Technical analysis of data can in turn push societal reflection on the goals and values we wish to pursue—large data sets enable auditing of institutions on a level that was previously

impossible, while AI systems create the potential to codify and operationalize ethical objectives at scale.

Foundational developments in the field of algorithmic fairness and accountability have come through works such as Datasheets for Datasets [32]—a generalizeable framework for documenting data sets. Datasheets are intended to be a standardized format for sharing fundamental information about data sets such as how the data was collected, funding sources, intended use, and how the data set will be updated. Datasheets are a simple but powerful tool that both encourage researchers to reflect on the data set creation and utilization process, while ensuring pertinent information is documented for future users. Other documentation frameworks—similar to (and often inspired by) datasheets—have been proposed such as model cards [33], network cards [34], Healthsheets [35], and interactive model cards [36]. While each of these approaches is has its own focus, they are unified in their effort to document the salient features of models and data sets in consistent, useful, and transparent fashions. These tools help work towards establishing norms in the field and provide the necessary information for broad sets of stakeholders to at least begin evaluating AI systems (e.g., returning to NIST's framing of "whether it does what is claimed"). With the rate of AI research rapidly increasingly [37] and the proliferation of AI technology in everyday life, having effective documentation procedures and norms are crucial underpinnings of fairness work.

There are broad categorizations of bias which can be helpful when analyzing a system—here we use the framework of Mehrabi *et al.* (2019) [38]. These biases include data set bias, algorithm bias, and human biases (NIST [31] presents a similar framing when identifying opportunities for bias mitigation). Data set bias in-

cludes measurement bias, where flawed measurements may be collected; representation bias, where sampling issues create non-representative data sets; and aggregation bias, where faulty assumptions lead to deriving insights about individuals based on membership in poorly constructed groups. Algorithm bias can result from decisions made when training (e.g., optimization functions), how users interact with an algorithm, or faulty evaluation approaches (e.g., biased benchmarks). Finally, human biases include historical bias, where existing societal bias permeates the data used to train AI; and social biases, where humans may feel social pressure to behave in a certain way when conducting surveys (or constructing AI systems).

Part of the dialogue opened up by work in the FATE space is a conversation around clearly defining what ethical objectives we wish to see in our society. At the end of the day, computers do exactly what we tell them to do. Even given stochastic elements or opaque processes, the user never-the-less issues some set of instructions that the machine faithfully executes (whether those instructions are what the user thought they were is another matter). Thinking about AI through the lens of FATE forces us to more explicitly say what we are optimizing for from an ethical perspective. In this dissertation we do not state what those broader ethical stances should be, but we do use some framings of fairness for the work on gender bias. General framings of fairness in the FATE space include ideas such as demographic parity, counterfactual fairness [39], and fairness through unawareness [40]. These three related forms of fairness motivate the work on patient health records and worker resumes. In the former case, we seek to develop a system where AI could be unaware of key demographic information when appropriate. In the latter case of resumes, our

work is partially motivated by the premise of demographic fairness—that is to say the gender pay gap should not exist.

Interpretable machine learning

Interpretable machine learning (IML) does not necessarily sit within FATE, but the concept is highly complementary to evaluations of fairness and bias. At the highest level, IML is an effort to provide insights into how machine learning algorithms make predictions. This can be accomplished by choosing more interpretable models (e.g., tree-based models, linear regression) or by applying specific IML frameworks to blackbox models. With linear regressions, one can look at feature importance within models to investigate how a model is making a prediction. For more complex models, it may be difficult to examine model internals in a meaningful way, so instead data perturbation is used to evaluate how varying inputs affect output [41]. There have been growing efforts to better understand at least some internal components of large lanugage models. For instance, BERTology has been coined as a body of work seeking to make BERT-based language models more interpretable. From these efforts, researchers have gained insights on how knowledge is represented within the BERT model architecture [42]. There are now full fledged libraries for IML analyses on even large, black box models—including tools for determining what pieces of a text language models are focusing on for a given output [43].

Demographic bias in language models

Language models encode the relationships present in their training data which includes biases along racial and gender lines. Seminal work on bias in language models

includes quantifying the gender dimensions of word-embeddings [44, 45] and attempting to debias word-embeddings from trained models [46]. Others have proposed modified training regimens to train language models to be less biased [47]. These types of debiasing efforts have been criticised for only addressing limited dimensions of bias, leaving additional demographic signals and biases unmitigated [48]. A language model adapted for medical contexts has been found to contain harmful biases including performing worse for non-dominant gender and racial classes [49].

1.2.2 Instruments and data sets in computational social science

To understand the sociotechnical systems, it is helpful to have instruments that can detect phenomena at varying scales. In the digital ecosystem of social media, many platforms report versions of this implicitly or explicitly—trending terms and entities are reported or suggested to users. For research purposes we need a more holistic accounting and broader access to the full breadth of conversation on platforms. A high-level aggregate view provides both indication of what narratives have achieved the highest levels of attention for a given day (e.g., US presidents on election day), as well as an initial entry point for research on more nuanced topics that may require tailored research (e.g., conspiracy theories that are simmering beneath the surface).

Storywrangler [50] provides an example of an instrument—derived from social media data—that makes an effort to begin to disentangle communications based on the most salient features. Other efforts blur the lines between instruments and data sets, but they share the goal of providing a view into sociotechnical systems.

Other examples include Hedonometer [51] for sentiment on Twitter, PushShift [52] for Reddit and other platforms, MediaCloud for conventional news media [53], and Google Ngrams [54] for books. We have seen important caveats for these tools, such as the Google Ngrams corpus composition changing over time and researchers not properly interpreting the cultural significance of results [55]. Examples of concerns for Twitter include the representativeness of the platform's users [56], the ability for researchers to capture the full breadth of user behavior [57], and the randomness of the Twitter API [58].

Storywrangler makes a few important distinctions for data during processing. The instrument ultimately serves n-gram time series, but classifies activities before doing so. The processing pipeline classifies communications based on the originality of the message (i.e., original tweets or retweets) and the language spoken in the message. The classification is an example of starting to peel back the layers of digital trace data and providing context for the final output.

Deliberate data

As we have seen with the training data for large-language models, the curation of data sets is often driven by technical objectives (e.g., sufficient data volumes for model size) and overt social concerns (e.g., the presence of blatantly offensive language). The same could be said for data sets used for algorithmic decision making in other areas. We need to think more about the data we are feeding to large language models. The current practice—defined by leading researchers as "feed the model with as much data as possible and minimal selection within these datasets" [11]—results in biased and often toxic behaviors in models. Bender and Gebru *et al.,* write "large, uncurated,

Internet-based datasets encode the dominant/hegemonic view, which further harms people at the margins" [59]. Even efforts to reduce bias and toxicity have actually increased these traits when carried out through simple target-word exclusion criteria [60].

Data is often not interrogated for demographic biases—issues such as who is covered in the data and how they are covered. Addressing these biases is crucial to fully understanding how the derived AI systems will behave, but also presents and opportunity to look at the underlying social system with a critical eye. As we will show in Chpt. 4, surfacing these biases we can both better understand the sociotechnical system as well as the AI systems we operationalize on top of them.

1.3 Ethics statement

There are a few ethical considerations we wish to highlight in the current studies. These can broadly be characterized as relating to the data sets we used, the framing of our research questions, and the broader impact of the results.

For the data sets, there are issues of privacy and representation. In the case of Twitter data used in Chpt. 2, individual tweets are only presented for public figures (i.e., US presidents) and data from other users' individual tweets is aggregated to the point of obscuring any individual contribution. There are concerns about sample bias and questions about who actually uses Twitter, with some studies finding Twitter to be younger and better educated than the US population in general [61]. For the electronic health records used in Chpt. 3 the data had been anonymized prior to our access (the records are housed at a research initiative that requires training and strict

usage agreements before granting access to bona fide researchers). The data was from a single hospital, presenting potential biases with the demographics of patients likely to access care in that setting. The resume data studied in Chpt. 4 was anonymized and accessed through a strict usage agreement with the vendor and was stored on a secure compute cluster. The representativeness of the resume data could not be established beyond gender, approximate age, and geographic location.

The framing of research questions for the study in Chpt. 2 was partially motivated by studying reactions from a broad base of individuals—however, owing to the biases in Twitter user demographics and those likely to engage with the tweets of US presidents we must acknowledge that many voices are left out. Indeed, this is but one instrument for sensing public response to communications and we should make an effort to listen to an array of sources. The last two studies (Chpts. 3 and 4) examine gender bias as the main point of inquiry. In both cases, we are limited to binary cases of gender due to the available data sets. The irony is not lost that in these studies, the additional dimension of analysis is in itself reduced to dipoles which are reductionist and miss broader insights.

Chapter 2

Ratioing the President: An exploration of public engagement with Obama and Trump on Twitter

2.1 Abstract

The past decade has witnessed a marked increase in the use of social media by politicians, most notably exemplified by the 45th President of the United States (POTUS), Donald Trump. On Twitter, POTUS messages consistently attract high levels of engagement as measured by likes, retweets, and replies. Here, we quantify the balance of these activities, also known as "ratios", and study their dynamics as a proxy for collective political engagement in response to presidential communications. We find that raw activity counts increase during the period leading up to the 2016 election, accompanied by a regime change in the ratio of retweets-to-replies connected to the

transition between campaigning and governing. For the Trump account, we find words related to fake news and the Mueller inquiry are more common in tweets with a high number of replies relative to retweets. Finally, we find that Barack Obama consistently received a higher retweet-to-reply ratio than Donald Trump. These results suggest Trump's Twitter posts are more often controversial and subject to enduring engagement as a given news cycle unfolds.

2.2 Introduction

The ability for US presidents to communicate directly with the public changed dramatically during the 20th and early 21st centuries. Moving from Franklin Delano Roosevelt's famous fireside chats through the television addresses of Harry Truman and John F. Kennedy, we now find ourselves in the era of social media campaigns and presidencies [62]. Communication technology has especially accelerated the ability for presidents to "go public" [63] and make appeals directly and instantly to the electorate.

With its introduction in the 2000s, social media provided a new platform for direct communication. New mechanisms for information sharing have consequences for contagion: rapid spreading of content through a user-base capable of responding in real-time. While these platforms democratize sharing for many categories of individuals on social media (e.g., celebrities interacting with fans), US presidential accounts present a unique opportunity to analyze particularly salient signals indicative of broader sociotechnical phenomena in an increasingly prominent aspect of politics.

Advertisers and political campaigns alike are concerned with engagement metrics for social media posts. Twitter itself markets its data as an early warning system for customer satisfaction and reputation management [64]—although the company has recently banned political advertising [65]. The rate with which a given post garners interactions (e.g., clicks, profile views, mouse-hovers, etc.) is a common measure of engagement in the digital realm. Activities in response to social media posts include retweets/shares, likes/favorites, and replies/comments. These activities all reflect specific user actions such as endorsing a post or expressing a divergent opinion. Some have theorized that simple ratios of activities may be used as proxies for of the polarity of the public's response to a given message [66, 67].

The term 'ratioed' is a Merriam-Webster "word we're watching" [68]. In the most general sense, the term refers to a social media post that receives more replies or responses than likes, favorites, or shares. In recent years the ratio has attracted growing interest by tech-journals and internet researchers, although this has largely been relegated to news articles and industry blog posts. [69–71]. On Twitter, the ratio value is generally taken to be defined by the ratio or balance of replies to likes or retweets. Here we will focus on the ratio of retweets to replies, as we show that like volume is often stable, while replies and retweets seem more reflective of specific public responses (Sec. 2.4).

In an effort to explore the dynamics of ratio space, in 2017 the website FiveThirtyEight Politics presented ternary ratios with activity counts normalized across retweets, comments, and likes [72]. Using this metric, the authors compared US politicians based on the ratio values their tweets receive. This work found noteworthy differences between politicians and political parties. As of late 2017, Trump tweets tended

to be met with relatively more replies, while Obama tweets were met with relatively more retweets. Tweets for both presidents had a high value of normalized activities that were likes (formerly known as 'favorites'). Beyond presidents, FiveThirtyEight Politics compared responses to the tweets of Republican and Democratic senators. The senatorial accounts exhibited normalized response values largely reflective of their party's most recent president—Republican senators tended to have high values of normalized reply activities while Democratic senators had more retweets and likes. We have found minimal academic research on the topic of response activity ratios and their temporal dynamics, although response activities have been studied separately.

More broadly speaking, time series of Twitter response activities have been used to investigate the relationship between on- and off-line political activity by analyzing correlations between current events and trends in cumulative activity sums [73]. Fundamental models relating retweet activity to external activity have been proposed at the hour and day scale for the 2012 South Korean presidential election [74]. Kobayashi and Lambiotte find future retweet activity to be log-linear correlated with early tweet levels [75]. Outside of Twitter, analysis of Instagram posts immediately preceding the 2016 US presidential election showed a higher volume of user activity for Trump supporters as well as more intense pro-Trump activity at the sub-day time-scale when compared to pro-Clinton posts [76].

Twitter specifically has been identified as playing an active role in the political arena as evidenced by studies of improved social organization during the Arab Spring [77,78], disinformation campaigns during the 2016 US presidential election [79–81], and political cohesion on the platform [82, 83]. Twitter has been shown to be a valuable source for data pertaining to domestic protests and social movements in

the US [84–86]. The 2016 US presidential election in particular has been extensively analyzed through the lens of Twitter data [87–90]. There is also increasing academic interest in the unique communication style of President Trump on Twitter [91, 92].

Twitter has been used to study several large-scale socio-technical phenomena such as the stock market [93, 94] and medical research [95, 96]. Messages on the platform have been shown to have a relationship with presidential approval rating polls [97–102]. Wang *et al.* examine how likes on Twitter can identify the preferences of Trump-account followers [103]. Although noisy at times, Twitter data can provide valuable insights on how audiences and populations respond to current events and specific messaging. Information diffusion and cascades on Twitter and other social networks has been the topic of much research [104–108]. Beyond politics on Twitter, Candia *et al.* propose a bi-exponential model to describe collective attention for online videos and manuscripts [109]. Media coverage of major events has been found to decay on roughly a weekly cycle, and scaling superlinearly with the population size of the affected area [110]. Amati *et al.* model retweet actions on dynamic and cumulative activity networks [111] while ten Thij *et al.* identify retweet graph characteristics common among viral content [112].

Some researchers have proposed critical levels of activity required to trigger information cascades on Twitter in light of tweet characteristics [113]. Others have proposed more general cascade requirements on social networks [114]. Jin *et al.* investigate how messages cross language and state communities on the platform [115]. Lee *et al.* predict retweet volume based on prior retweet spacing, user meta-data, tweet content [116]. Others have identified celebrity-users (including Barack Obama) that are central in spreading new content [117].

Regarding the representativeness of social media data, several studies have examined the degree to which Twitter represents the general population and voting public [61, 118–120]. Grinberg *et al.* investigate fake news exposure for Twitter accounts belonging to eligible voters and found disproportionate exposure to a small number of accounts [121].

How likely are Twitter users, and specifically POTUS account followers, to be eligible voters in the US? Pew Research Center has estimated that 26% and 19% of US adults on Twitter follow the Obama and Trump accounts, respectively [56]. The New York Times estimated however that less than 20% of Trump's followers are voting-age US citizens [122]. These considerations are important to remember when attempting to establish a relationship between Twitter metrics and political outcomes. Nevertheless, the platform has been shown to be a powerful tool for evaluating public sentiment towards politicians, and even detecting adversarial actions in the US political process.

Presidential communications are a timely topic in light of rapidly changing norms surrounding how the executive branch communicates with the public. Moreover, highly influential Twitter users such as US presidents produce signals in the medium that largely transcend the social network dynamics. When a president tweets, the message is nearly instantly amplified and spread by official and unofficial sources on the platform. Indeed, US presidents are often discussed at rates approaching function words on Twitter (i.e., the rate of usage of the name "Trump" may approach usage rates for words such as "the" and "RT" on the platform) [123]. Examining ultrafamous users somewhat removes the more prominent effects of network topology

on the response activity and provides an opportunity to view response behavior largely as a function of timing and message content.

The volume of response activities on Twitter is arguably the most tangible metric that is publicly available for evaluating user-base responses to content on the platform. Furthermore, these values are highly visible to users and may influence behavior when users seek to affect the collective response to a communication (e.g., 'take a stand' by responding [124]). Establishing characteristic scales of activity volume and temporal dynamics for engagement is an important step in understanding how these values can provide insights on user-base response.

Counts of user activities responding to tweets can be readily viewed through the icons at the bottom of every tweet, which also provide the interface for user engagement. These values provide the end user with an indication of the tweet popularity and, in some cases, controversiality. Taken together with the cultural context surrounding the original tweet, the ratio of these values open up the possibility of studying tweets through the lens of "ratiometrics". This allows for the distillation of response activities into aggregated measures of the user base's reaction. Moreover, enables the comparison of response activities on Twitter across accounts and across time. From there, we can begin to use the ratio values as a criteria when evaluating the content of tweets.

A simple calculation of the publicly accessible response activity counts is a starting point, but insufficient to fully investigate the full potential of ratiometrics. Here we present a suite of tools—collectively referred to as the "ratiometer"—that help us understand response activity profiles for tweets. Many factors contribute to interpreting the activity counts for a tweet, including typical activity ratios for the user and

the age of the tweet. Understanding the typical response that an account receives requires building a historical view of the user's tweets and their subsequent response activities. Another challenge is determining the typical response volume at a given time step since the tweet was issued.

We show example time series for retweets and replies in response to four Trump tweets authored after the 2016 election in Fig. 2.1. In the case of an ultra-famous user like Trump, we see an immediate response to tweets with early retweets and replies occurring seconds after the original tweet. For Trump's tweets from after the 2016 election, we generally observe thousands of response activities within the quarter-hour, with response-activity counts nearing their final values within 24-hours of the original tweet. There may be modest growth during the proceeding week, but generally these values have stabilization periods at the day-scale (S1 Fig. and S2 Fig.). The insets in Fig. 2.1 show the retweet-to-reply ratio time series. These values often fluctuate even while individual ratio time series appear to have a more stable trend. The above highlights some of the challenges in building an understanding for the expected volume and time scale associated with response activities for a specific account.

In the present study, we explore the time series of the volume of user activities in response to presidential tweets. We summarize the ratio values of responses to Obama and Trump tweets over three distinct periods of recent US political history. Our investigation includes ratio values for all three activity counts normalized against the sum of all activities (we refer to these as ternary ratios below). We also present characteristic time scales for response activity volume over the three political periods for both presidents. Finally, we present words that tend to appear more often in

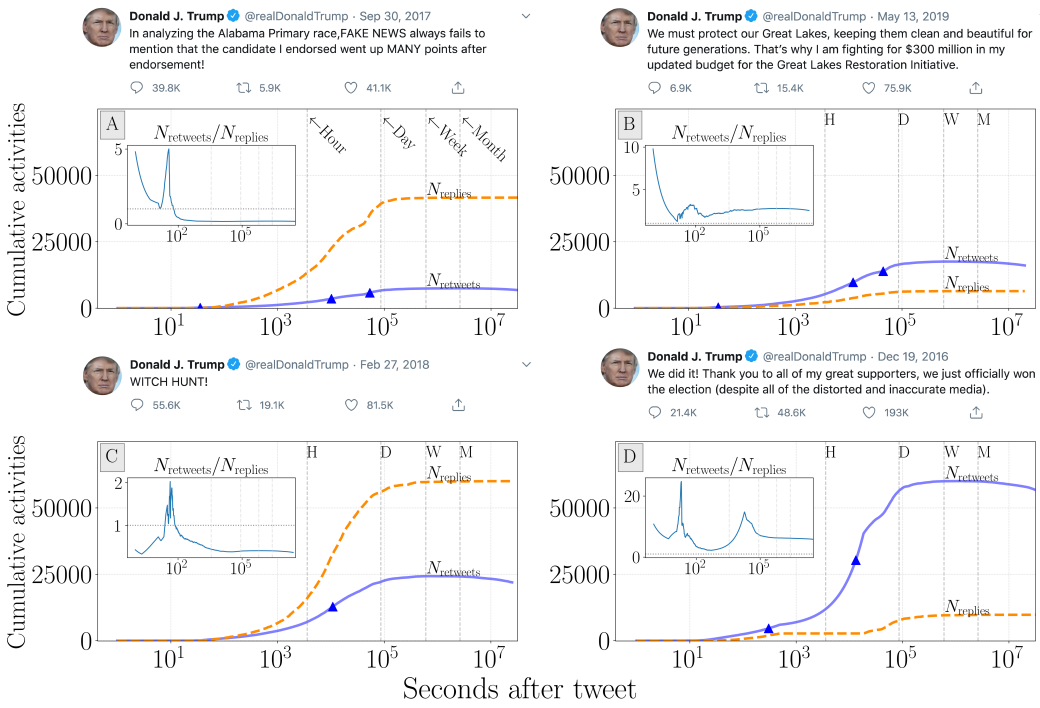


Figure 2.1: **Time series for the cumulative sum of retweet and reply activities** *for four tweets authored by the Trump account after the 2016 presidential election. The temporal axis is logarithmically spaced to show early-stage growth. Insets represent the cumulative ratio of* $N_{\text{retweets}}/N_{\text{replies}}$ *for the same time period. Vertical lines with H, D, W, M correspond to hour, day, week, and month intervals. Triangles indicate inflection points where the second derivative of retweet volume transitions from positive to negative (see Sec. 2.4.2 and Fig. 2.6).* N_{retweets} *can decrease because the values are retrieved directly from the Twitter data—with account deletions leading to declines over time. The tweets examined here only provide some illustrative examples of how response activity time series behave.*

ratioed tweets for the @realDonaldTrump account. In section 2.3.1, we describe our data and methodology. In section 2.4, we present and discuss the results of our study. Finally, in section 2.5, we offer concluding remarks and point to areas for future work.

2.3 Methods

2.3.1 Data

We construct the dataset analyzed in the present study by filtering Twitter's Decahose Streaming API [125] an, in principle, random 10% sample of tweets authored since 2009. The stream contains a variety of activity types including tweets, retweets, quote tweets, and replies. While the Twitter REST API serves up-to-date information for a given user or activity, the Streaming API data offers a snapshot of each activity at the moment of generation. The specific form of this dataset allows us to create a historical timeline of activities; responses to each original activity arrive with timestamped counts for several metrics including replies, likes, and retweets. For the purposes of this study, we define 'activity' to refer to any user action recorded in the historical sample including original tweets, retweets, retweets with comments, and replies.

Given the sampling rate of the Decahose API, each individual activity has an approximately 10% chance of appearing in our data set. For popular users and tweets (i.e., tweets that garner retweets and replies that number in the thousands or more), we are almost certain to record activities responding to the original activity. Moreover, we are able to observe a number of activities with sufficient temporal resolution to construct historical time series down to the level of a single second.

For this study, we are looking at 2 very popular users on Twitter. Their tweets regularly receive sufficient response-activity volume for our sampling method to be effective. It is worth noting that our method would yield noisier estimates for less popular accounts (and tweets).

2.3.2 Tracking Retweets and Replies

Starting with the 10% random sample of Twitter activities, we collect activities responding to presidential tweets by filtering for replies and retweets responding to the @BarackObama and @realDonaldTrump accounts. We do not include the official US Presidential Twitter account (@POTUS) or the official White House account (@WhiteHouse). The random sample feed serves original tweets, retweets, and replies. Each of these activities are served in a tweet object that contains metadata which includes creation time and follower counts for the author of the activity. In the case of retweets, the tweet object contains a count of retweets and likes at the moment the activity is retweeted. Replies are tagged with user metadata and time of activity. Altogether, the metadata provides the historical data allowing construction of the response activity timeline.

Quote tweets appear in our data stream but we do not count them towards our tally of retweets. This is partially motivated by a desire to maintain backwards compatibility of our analysis with earlier periods before quote tweets were an official feature on Twitter. In a preliminary analysis of 20 Trump tweets with high engagement, we determined that for 16 out of 20 tweets the number of replies and quote tweets both out numbered the number of regular retweets. We mention this as reference point but leave in-depth analysis of quote tweets to future research.

2.3.3 Measuring the Ratio

While retweets counts are observable via tweet metadata, replies must be gathered by maintaining a cumulative summation of reply activity counts (multiplying the final value by 10 to account for our sample rate). We collect counts of retweets and likes from the metadata, while replies are tallied using the cumulative sum method. Validating these estimates for final reply values, we find the error to be less than 1% in terms of actual/predicted reply counts for high activity tweets (actual reply counts were collected for a random sample of tweets via the Twitter web-interface).

In order to calculate the ratio at a common time step, we linearly interpolate the values for replies, retweets, and favorites, resulting in time series that can be sampled at arbitrary intervals for all activity types. We report time steps at the second resolution unless otherwise stated.

There are several challenges associated with the historical feed, stemming from contradicting metadata on activity objects that occurred simultaneously. With timestamps at the 1 second resolution, sometimes tweet activities are reported as occurring at the same time (inferring time stamps from Twitter snowflake IDs did not resolve these conflicts [126]). Activities that reportedly occur at the same time often have conflicting values for retweets and likes (sometimes differing by hundreds of activities). This may be due the count values growing so quickly that the activities within a given second window have differing values and/or it may be an artifact of the slow update times within the platform's database. This latter point is made more notable when taken in combination with Twitter's practice of deleting millions of tweets per week (this would lead to fluctuations in the activity counts for a given tweet).

Regardless of the cause, fluctuations in activity counts result in time series that increase non-monotonically. To make model fitting and data analysis easier, we remove observations that result in non-monotonically increasing behavior before the maximum value of the activity time series. When calculating first and second derivatives we use a Gaussian filter to smooth the time series so that we avoid incorrectly identifying noisy regions as notable transitions from positive to negative derivatives. There is decay in values for retweets while replies only increase—the exact reasons for the decay are not investigated here, but the lack of decay in the replies is due to the cumulative sum method we use for estimating reply values. There is further work to be conducted on the possibility that these jagged portions are a signal of banned account activity, and the dynamics of the jagged regions may indicate activity by accounts that the platform ultimately deems ban-worthy.

Ternary activity values are calculated by dividing each activity count by the sum of all activities at a given time step. The ternary ratio value for activity type τ at time step t is given by

$$\mathcal{R}_\tau(t) = \frac{N_\tau(t)}{N_{\text{retweets}}(t) + N_{\text{likes}}(t) + N_{\text{replies}}(t)}\,, \tag{2.1}$$

where $N_\tau(t)$ is the count of the activity at time step t. With the above each observation is comprised of a 3-dimensional vector representing the normalized activity values for a tweet.

The ternary ratio values are represented on a ternary plot (2-dimensional simplex) where the values of activities at each time step sum to 1,

$$\sum_\tau \mathcal{R}_\tau(t) = 1\,. \tag{2.2}$$

2.3.4 Vocabulary of Ratioed Tweets

We are also interested in how the text content of tweets relates to ratios of response activities. To investigate this we calculate rank turbulence divergence—as defined by Dodds *et al.*—between ratioed and non-ratioed tweets [2].

The rank turbulence divergence between two sets, Ω_1 and Ω_2, is calculated as follows,

$$\begin{aligned} D_{\alpha}^{R}(\Omega_1||\Omega_2) &= \sum \delta D_{\alpha,\tau}^{R} \\ &= \frac{\alpha+1}{\alpha} \sum_{\tau} \left| \frac{1}{r_{\tau,1}^{\alpha}} - \frac{1}{r_{\tau,2}^{\alpha}} \right|^{1/(\alpha+1)} , \end{aligned} \tag{2.3}$$

where $r_{\tau,s}$ is the rank of element τ (n-grams in our case) in system s and α is a tunable parameter that affects the impact of starting and ending ranks.

While there are numerous techniques to compare two corpora, we use rank divergence in light of its robustness when working with subsamples of heavy tailed distributions. Working with ranked values also alleviates some challenges that may arise when normalizing n-gram occurrences to arrive at frequency of usage. Further, rank divergence gracefully deals with the problem of n-grams only occurring in one of the two corpora. Finally, ranked values make interpretation more straightforward when presenting results (with the reader need not worry about normalization, etc.). These factors combined with the tunability of rank divergence lead us to find it was best suited for our use case.

Here, we take the content of tweets from the Trump account after his declaration of candidacy on June 16, 2015. We then split this set by ratioed ($N_{\text{retweets}}/N_{\text{replies}} < 1$) and non-ratioed tweets ($N_{\text{retweets}}/N_{\text{replies}} > 1$). Obama's account did not have a

sufficient number of ratioed tweets to conduct this analysis. From here we calculate the rank divergence, between the two sets. We set $\alpha = 1/3$ based on experimentation outlined in the original rank divergence piece by Dodds *et al.* This value of alpha tends to balance the influence of high- and low-ranked items.

2.4 Results

2.4.1 Overall time series

Viewing replies, retweets, and likes for individual tweets over time, we see how activities in response to Trump and Obama tweets have changed in the years around the 2016 election. In Fig. 2.2, we show the final count (the activity volume 168 hours after the original tweet is authored) of activities for each tweet in the POTUS data set. Then-candidate Trump's tweets experienced a notable increase in response activity following his June 16, 2015 declaration of candidacy. Prior to this point, Trump's tweets garnered 2 to 3 orders of magnitude less activities than during the height of his campaign and his subsequent time in office.

Because of the difficulty in sampling replies using our data (see Sec. 2.3.1 for tweets that receive few replies), many of the tweets from the Trump account from before his candidacy have a high degree of uncertainty in their true value. This uncertainty is introduced by our method for inferring reply counts introduces variance in the overall activity balance.

In Fig. 2.3 we present the normalized activity values for tweets from Obama and Trump from 2013 to 2019. The ternary histograms allow us to compare all activity

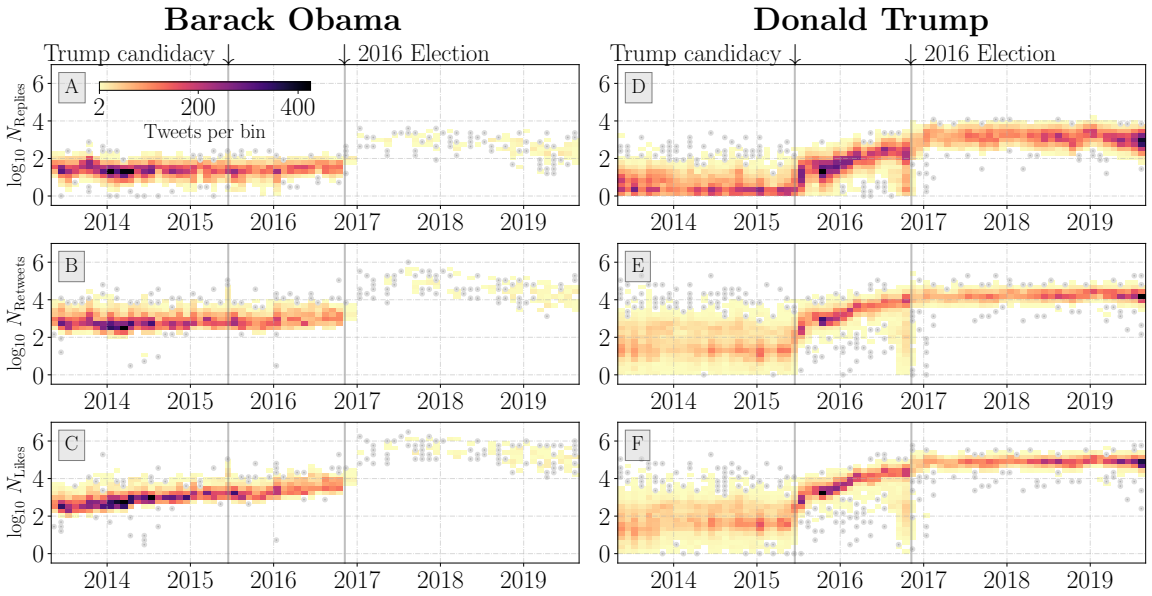


Figure 2.2: **Time series histogram of maximum activity counts for tweets** *posted from the Barack Obama (**A**–**C**) and Donald Trump (**D**–**F**) Twitter accounts. Each bin represents a collection of tweets at their original author date along with their respective maximum observed activity count. Bins with less than 2 tweets are shown as grey dots. We include all tweet types (e.g., advertisements, promoted, etc.); see Section 2.3.1 for information on collection methods. We annotate Trump's declaration of candidacy and the 2016 US general election with solid vertical grey bars. A marked decrease in Obama account activity is apparent immediately following the 2016 election. The region of outliers in the Trump time series immediately preceding the 2016 election has been determined to be largely reflective of promoted tweets which have abnormal circulation dynamics on the platform [1].*

values, while the time series in Fig. 2.3D and Fig. 2.3H show the retweet-to-reply ratio. We choose the latter as it appears to offer a more descriptive measure of response activities (i.e., counts of likes are generally more stable within sub-regions of the yearly time series). The Obama $N_{\text{retweets}}/N_{\text{replies}}$ time series (Fig. 2.3D) demonstrates how the Obama account tends to receive more retweets than replies (median ratio value of 3.67 between June 6, 2015 and November 8, 2016).

Once Obama leaves office, the account receives consistently more retweets than replies on the limited number of tweets authored (median ratio value of 7.77 after November 8, 2016).

The Trump $N_{\text{retweets}}/N_{\text{replies}}$ time series (Fig. 2.3H) shows the tendency of then-candidate Trump's account to be ratioed less during the campaign season. Soon after transitioning to office the Trump account begins receiving more replies relative to retweets—with a transition period roughly corresponding to the time between the day of the election and inauguration. For the campaign period Trump has a median ratio value of 3.1 while after the election the account garners a median ratio value of 1.32.

In Fig. 2.3E we show a ternary histogram for the pre-campaign Trump account, where the high variance in reply estimates are evidenced by broad coverage of most regions of the simplex. Error in estimating replies alone does not account for this high variance, with the like and retweet time series from Fig. 2.2 showing higher variance for Trump as well. Another time series artifact is the presence of Trump tweets that were met with low activity counts around the 2016 election. Around the same time, the Obama account activity drops precipitously. After the election, Obama's limited number of tweets are met with consistently high counts of likes, retweets, and replies.

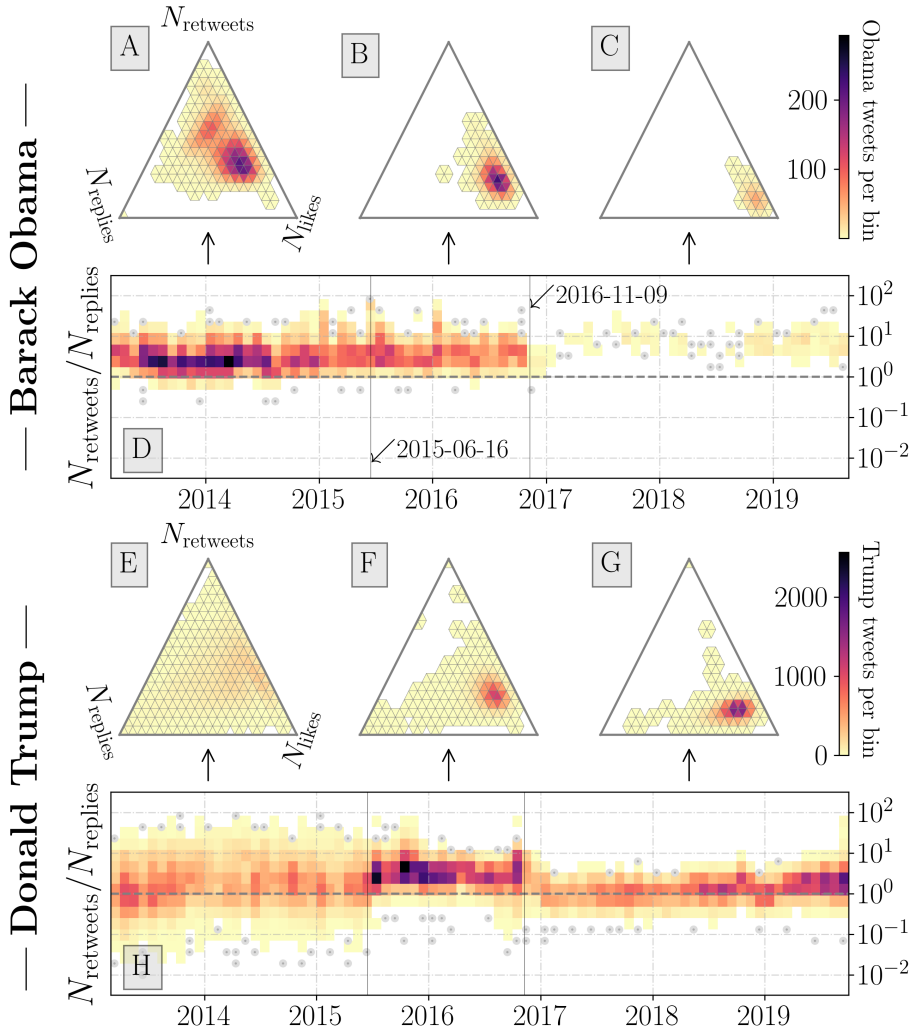


Figure 2.3: **Ternary histograms and $N_{\text{retweets}}/N_{\text{replies}}$ ratio time series** *for the @BarackObama (**A**–**D**) and @realDonaldTrump (**E**–**H**) Twitter accounts. The ternary histograms (**A**–**C** and **E**–**H**) represent the count of retweet, favorite, and reply activities normalized by the sum of all activities. White regions indicate no observations over the given time period. See Fig. 2.4 for examples of full time series for response activity for example tweets. Heatmap time series (**D** and **H**) consist of monthly bins representing the density of tweets with a given ratio value. Single observations (bin counts < 2) are represented by grey points. The two dates annotated correspond to the date of Trump's declaration of candidacy ($2015-05-16$) and the 2016 general election ($2016-11-09$). We show the tendency for Trump tweets to have ternary ratio values with a greater reply component—with pre-candidacy tweets having higher variability and pre-election tweets having a higher $N_{\text{retweets}}/N_{\text{replies}}$ ratio value. Post-election Obama tweets have ternary ratio values with more likes than other periods for both Obama and Trump.*

The rapid increase in activities in response to Trump tweets, and the corresponding decrease in the overall variance of counts for activities, are important insights visible in Fig. 2.2. These measures are indicative of the meteoric rise of then-candidate Trump, along with his now pre-eminent Twitter presence (in 2020 his name appears more frequently than the word "god" on most days [123]). While the two time periods are not directly comparable, by the end of the 2016 election, Trump's account consistently garnered more response activities than President Obama's. After the 2016 election, the Obama account's tweeting frequency is reduced while also experiencing a notable rise in response activities. These results helped inform the selection of distinct periods around the 2016 election. These time periods were both meaningful in a political context—marking Trump's declaration of candidacy and the 2016 election day—as well as in the context of response activity time series.

2.4.2 Example Ternary Time Series

Using observations of responses to historical tweets, we can construct the time series for all ratios of activities. In Fig. 2.4 we show the ratio time series for three Trump tweets. We selected the tweets based on the value of their final retweet-to-reply ratio—with a tweet from approximately the 90th, 50th, and 10th percentiles of final $N_{\text{retweets}}/N_{\text{replies}}$ ratio values. The response activities to most tweets (with high response activity) tend to experience some early volatility, partially owing to the low number of observations. One hour after the original tweet has been authored, the response ratios tend to fall into a stable region, or at least adopt a stable trend. Within this signal there is also the effect of bots (S2 Appendix) that are likely programmed to respond to Trump account activity within seconds.

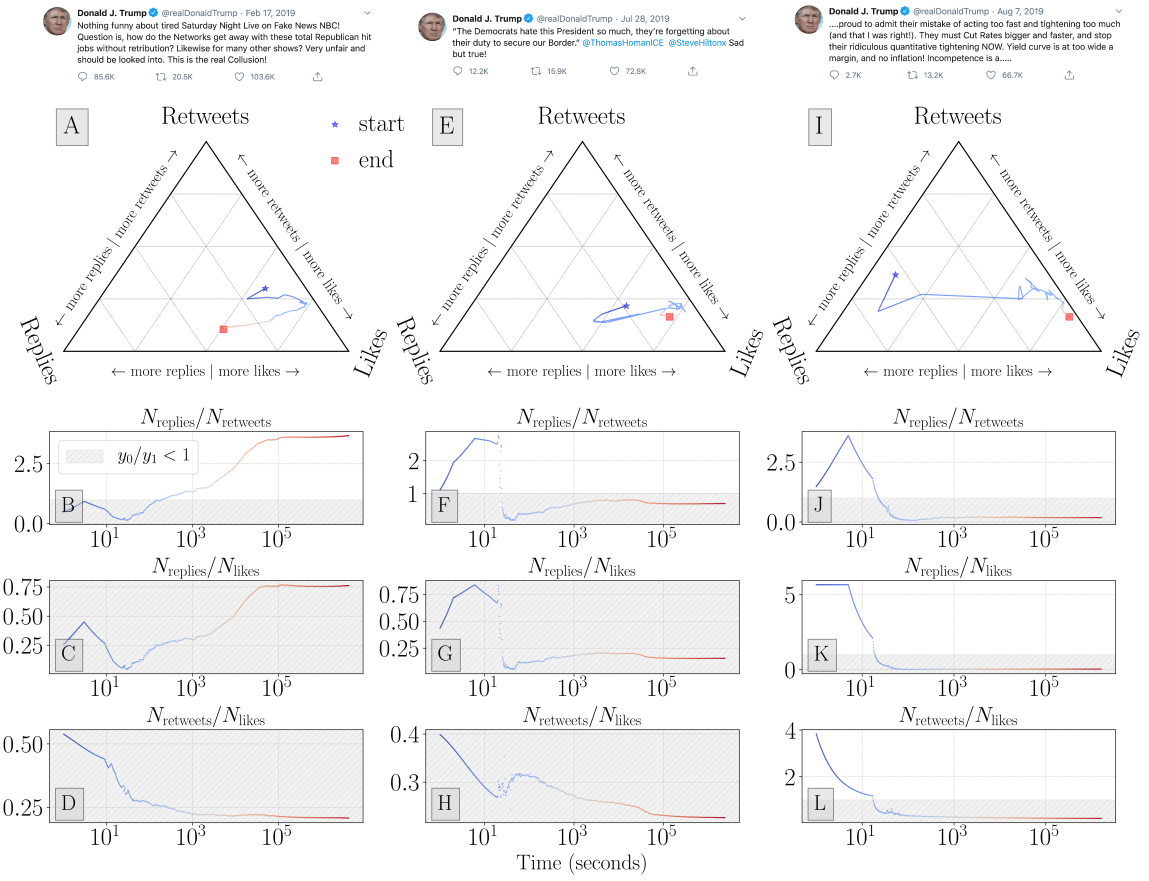


Figure 2.4: **Ternary time series for three popular Trump tweets**, *selected to represent messages in approximately the upper 90th percentile, 50th percentile, and bottom 10th percentile of* $N_{\text{retweets}}/N_{\text{replies}}$ *ratios. The time series represent observations from first activity observation to the final observation of the tweet. These ternary time series contrast the simple 2-dimensional ratio trajectories and illustrate example trajectories through the 3-dimensional ratio space. See S1 Fig. and S2 Fig. for the distribution of ternary ratio values for Obama and Trump tweets over time. Each of these three tweets were authored on May 25, 2019. Direct links to tweets for Screenshots were collected on May 28, 2020.*

We show how ratios tend to stabilize by presenting ternary histograms of activity response ratios over the seconds and days after a tweet is published. S1 Fig. shows how the Obama account largely has ratios biased towards retweets and likes throughout the period after a tweet is authored. Ratios for the Trump account (S2 Fig.) tend to have greater variance in their ratio values in the seconds and hours after a tweet is released. The ratios for Trump's account are also biased towards more replies, relative to the Obama account, throughout the period after issuing a tweet. This is consistent with the final ratio values for each president across three political periods surrounding the 2016 election.

2.4.3 Distribution of Ratios

The final observed ratio value for a tweet offers an aggregate measure of user-base reactions. We examine tweets which were authored at least 168 hours prior to the activity calculation to ensure the responses have largely stabilized. Of course, there is still the possibility that users will respond to the tweet long after the original activity period, but we have found these lagging response activities to minimally affect the normalized ratio value. It is worth noting that this delayed response behavior is common for prominent users such as Obama and Trump, with some response activities taking place months or years after the original activity (often in reference to current events that are addressed in the old tweet: 'there is always a tweet').

In Fig. 2.5 we show the distribution of Trump and Obama $N_{\text{retweets}}/N_{\text{replies}}$ ratios, highlighting the tendency for the Trump account to "get ratioed" more often relative to the Obama account. Of 13,639 tweets in Fig. 2.5 from the Trump account, 18% receive more replies than retweets. The Obama account generates a relatively limited

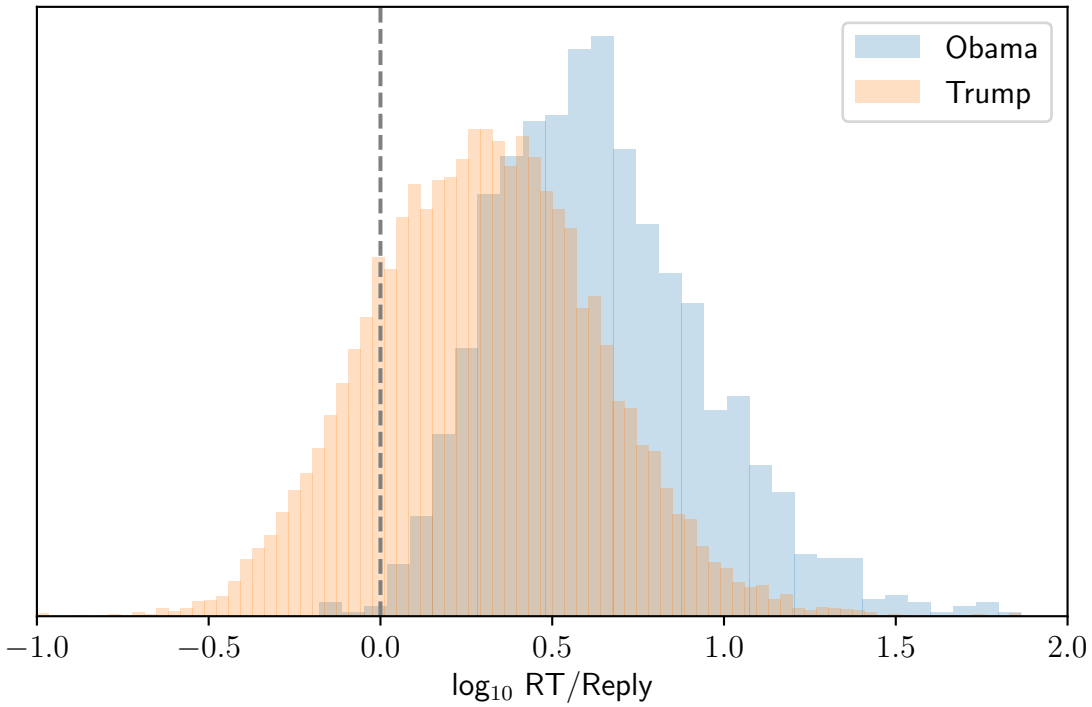


Figure 2.5: **Histogram of final observed ratios for tweets authored by @BarackObama and @realDonaldTrump accounts**. *Observations left of the dashed vertical line correspond to tweets that are considered ratioed. The shift in the distribution corresponding to Trump's account can be plainly seen—with the account producing more tweets that are ratioed than compared with Obama's account. For both accounts, tweets shown here are restricted to those authored after Trump's declaration of candidacy. Of 16,708 tweets included from the Trump account, 3,015 (18%) have a* $\log_{10} N_{\text{retweets}}/N_{\text{replies}}$ *value less than or equal to 0 and 13,693 (82%) have a score greater than 0. Of 1,786 tweets from the Obama account, 7 (*$< 1\%$ *have* $\log_{10} N_{\text{retweets}}/N_{\text{replies}}$ *values* ≤ 0 *while 1,779 (*$> 99\%$*) have values* > 0*. From the distribution of ratio values we see that ratioed tweets are outliers for the Obama account while the Trump account often gets ratioed and has a lower ratio value on average.*

number of tweets that receive more replies than retweets (less than 1%, see S4 Fig. for Obama's ratioed tweets). This simple ratio calculated with two of the three activity measures loses some context in terms of overall user-base responses, and we now move to incorporate the like activity volumes.

We subset the tweets for Presidents Obama and Trump based on the time periods introduced in Fig. 2.2. In Fig. 2.3 we show that for both Obama and Trump accounts there is a higher degree of variation in final ternary ratio values for earlier years. This is likely in part due to the reply thread mechanisms and structure changing on Twitter's platform, as well as the growing popularity of political accounts on Twitter, aside from background changes in how users engage on the platform. This shift is especially prominent for the Trump account, which experienced a marked growth in response activity after Trump's declaration of candidacy. The higher variance of final ratio values is accentuated when the sample of replies is lower, owing to our method of estimating reply activities. Comparing Obama and Trump ternary histograms (Fig. 2.3) we observe more Trump activity ratios in the reply region of the simplex.

For the Obama account, Figs. 2.3A–D, we see notably higher values for normalized retweet and like activity compared with the Trump account. The retweet-to-reply ratio is consistently higher in the time series presented in Fig. 2.3D. Once leaving office, the Obama account demonstrated a further tendency to receive a higher proportion of likes and retweets.

2.4.4 Characteristic Time Scale

To empirically investigate the characteristic time scale of response activities, we report the time index t for the local maxima of the instantaneous retweet count for each tweet (the local maxima of the first derivative of cumulative retweets N_{retweets}, or $\texttt{argrelmax}\,\frac{d}{dt}N_{\text{retweets}}$). More specifically, we use the second derivative test to find points where the second derivative of cumulative retweet counts drops below 0,

$$\frac{d^2}{dt^2} N_{\text{retweets}}(t_{i-1}) > 0 \text{ and } \frac{d^2}{dt^2} N_{\text{retweets}}(t_i) < 0\,. \tag{2.4}$$

These points are viewed as being indicative of the start of a decreasing trend for the volume of new response activity, specifically marking the moment when the rate at which new activities are generated begins decreasing. Going forward we will refer to these points as inflection points (examples of these points can be seen with triangular markers in Fig. 2.1).

A similar measure could examine the point where the second derivative rises above 0. This would be a valid measure, but we prefer our measure for a couple of reasons. First, we found it better characterizes the time scale of the initial burst of activity that most POTUS tweets receive (i.e., how long the first wave of response activities is sustained). Second, it provides a closing bound on a period of increased response activity (i.e., the point indicates when activity starts to fade). We were more interested in characterizing when attention, as indicated by response activity, is no longer maintained for the POTUS tweets. We include results for "inverse-inflection" points in S3 Fig. as a point of reference.

When viewing time since an original tweet in Fig. 2.6, we see a higher density of inflection points around the 1 minute and 1 day time intervals. This suggests that many periods of intense activity take place within the first day of a tweet being authored. We also observe inflection points in cases where more sporadic engagement occurs later in the tweet response-activity timeline. This leads to another spike in activity around 1 week after the original tweet. Over 90% of inflection points are observed before 10^5 seconds (~ 1.15 days) for both accounts and for all time periods as evidenced by complementary cumulative distribution functions (see Fig. 2.7).

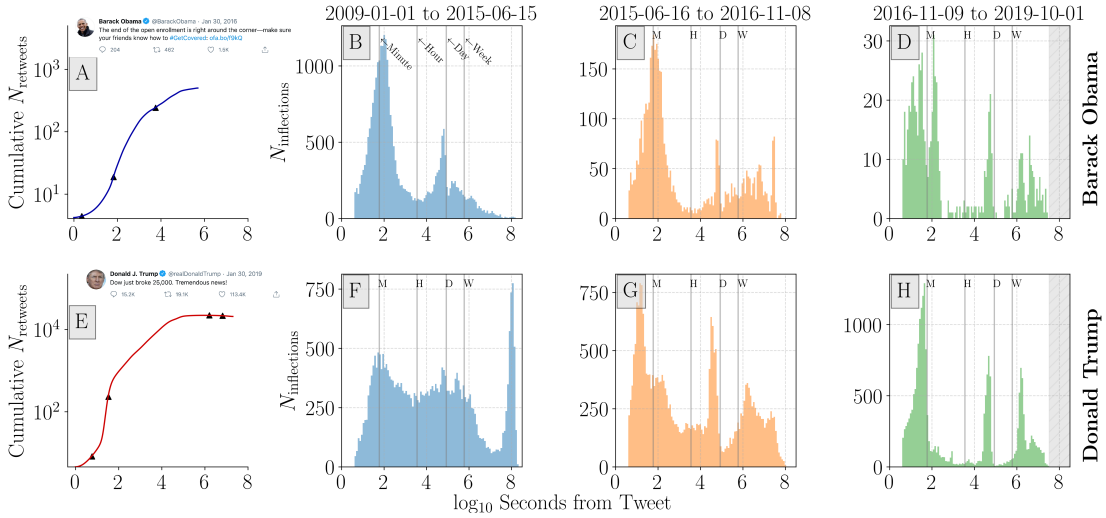


Figure 2.6: **Distribution of inflection points**, *the local maxima of instantaneous retweet volume,* `argrelmax` $\frac{d}{dt}N_{\text{retweets}}$, *where* $\frac{d^2}{dt^2}N_{\text{retweets}}(t_{i-1}) > 0$ and $\frac{d^2}{dt^2}N_{\text{retweets}}(t_i) < 0$. ***A*** *and* ***E****: Example cumulative retweet time series and inflection points (solid triangles) for Obama and Trump tweets. Histograms show the distribution of inflection points across all tweets binned by time periods before (****B*** *and* ***F****), during (****C*** *and* ***G****), and after (****D*** *and* ***H****) the 2016 US presidential election campaign. The January 1st 2009 to June 15th 2015 period for Trump (****F****) contains inflection point counts that are largely reflective of low initial activity and high(er) late activity (months or years later) leading to unusually high values for seconds to first inflection point (*$> 10^8$ *seconds). For Obama's and Trump's time in office, tweets experience inflection points around 1-minute and 1-day after the tweet is authored—indicating characteristic time-scales of activity waning. Histogram bin widths are in effect logarithmically spaced over the displayed time span. This has the effect of providing a higher temporal resolution for shorter timer periods. Screenshots were collected on May 28, 2020.*

There is a notable difference between the governing periods for Obama and Trump, Figs. 2.6A and 2.6G, respectively. Trump tweet response time lines demonstrate a tendency to generate inflection points before the 1-minute mark—indicative of rapid response to his account's tweets. Additionally, we find more time series inflection points after 1-week has passed for the Trump time series than compared with Obama. This suggests that users may return to Trump tweets later to comment on the content (perhaps as new political developments occur). This is further illustrated in Fig. 2.7B, where we show that 10% of Trump inflection points take place after 10^6 seconds ($\sim$ 10 days). This is true for the pre-campaign, campaigning, and governing periods. By contrast, 90% of Obama inflection points take place before 10^5 seconds ($\sim$ 1 day).

2.4.5 Ratio Content

We can investigate how tweet content relates to response activities by comparing the distributions of words that appear in ratioed and non-ratioed tweets. Taking the rank of the frequency of occurrence for each group, we then compare the rank-turbulence divergence for the two groups [2]. Per the allotaxonograph in Fig. 2.8, we are able to tune the impact of starting and ending rank magnitude on the overall score for the word. The allotaxonograph is an instrument that allows us to compare the Zipf distributions of two corpora. The figure features a rank-rank histogram and a vertical bar plot describing each 1-grams' contribution to the overall rank divergence value.

In Fig. 2.8, we show that ratioed tweets contain more words related to fake news, the Mueller inquiry into Russian interference, and Colin Kaepernick and kneeling during the national anthem in the National Football League (NFL). For non-ratioed tweets, we see words from campaign-related communications (e.g., "Jeb", "Iowa", and

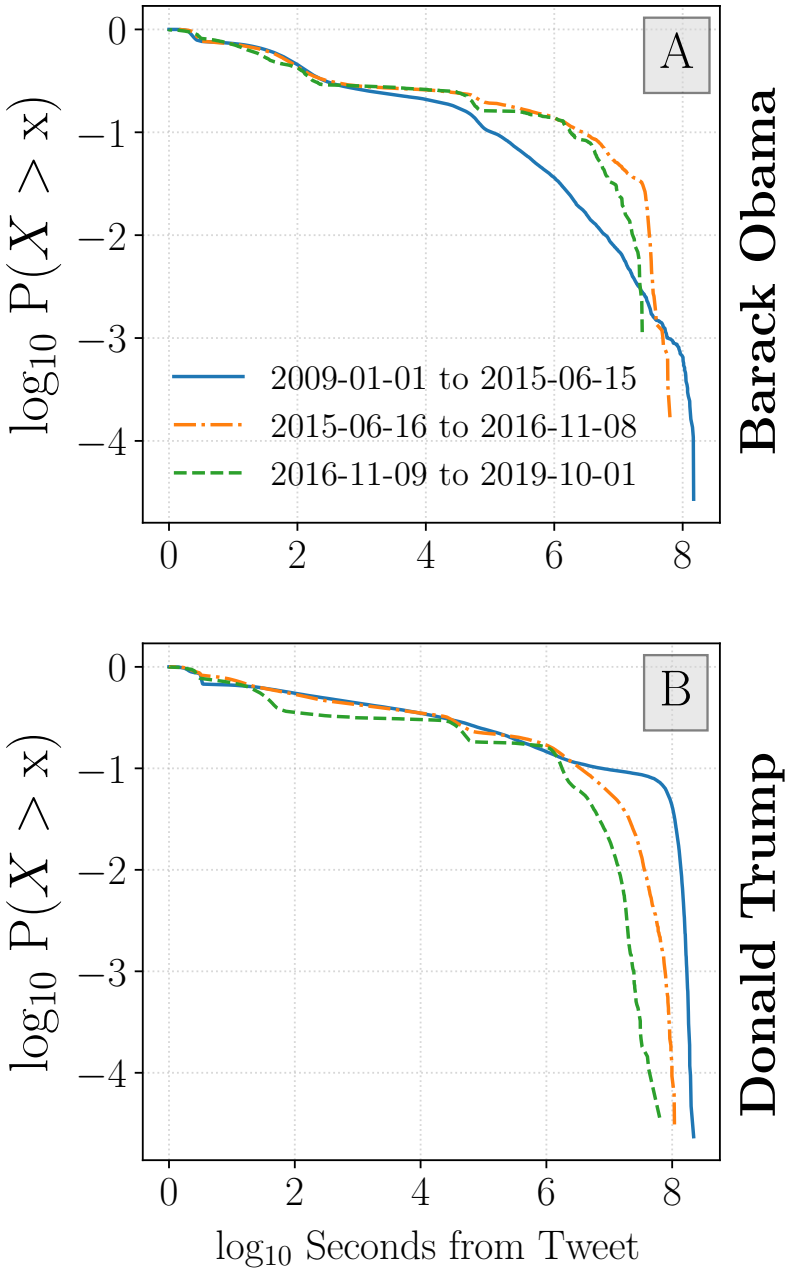


Figure 2.7: **Complementary cumulative distribution function for inflection point timing.** *Roughly 90% of Obama inflection points (**A**) took place before* 10^5 *seconds (*$\sim$ 1 *day) for the period before the 2016 election cycle. For the period during and after the 2016 campaign season, both Obama and Trump tweets (**B**) receive 10% of inflection points after roughly* 10^6 *seconds (*$\sim$ 10 *days) from the initial tweet.*

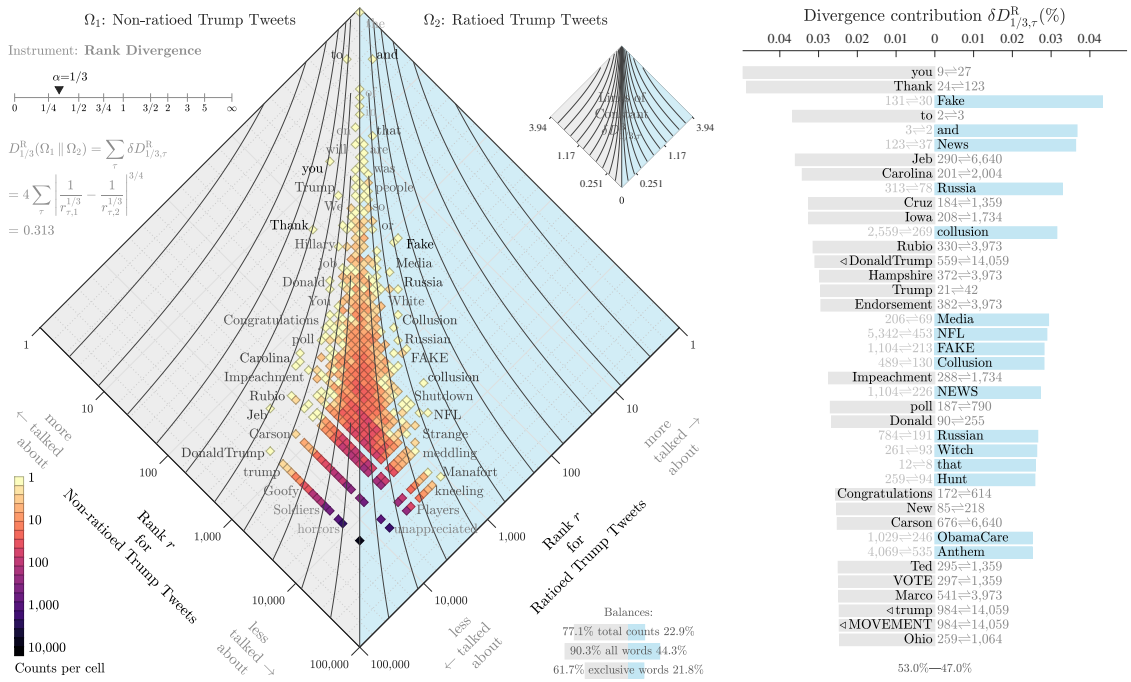


Figure 2.8: **Rank divergence allotaxonograph** *[2] for* 1*-grams from Trump account tweets authored after Trump's declaration of his candidacy on June 16, 2015. "Ratioed" tweets are those where the proportion of retweets over replies is less than 1 ($N_{\text{retweets}}/N_{\text{replies}} < 1$), non-ratioed tweets accumulated a ratio greater than 1 ($N_{\text{retweets}}/N_{\text{replies}} > 1$). There is a notable class imbalance. The majority of tweets are non-ratioed, with a median value of ~ 2. To generate the above figure we examined 3,313 ratioed tweets and 15,274 non-ratioed tweets. The* 1*-grams "Fake", "News", "Russia" and "NFL" are ranked higher in the ratioed corpus. For the non-ratioed corpus, the* 1*-grams "Jeb", "Carolina", and "Ted" are ranked higher. This illustrates the tendency of ratioed tweets from this period to contain more politically contentious* 1*-grams related to Trump scandals. Non-ratioed tweets from this period more often contain campaign related messages. In the main horizontal bar chart, the numbers next to the terms represent their rank in each corpus, while terms that appear in only one corpus are indicated with a rotated triangle. The smaller three vertical bars describe the balances between the ratioed and non-ratioed corpora: 77.1% of total counts occur in the non-ratioed corpus; we observe 90.3% of all* 1*-grams in the non-ratioed corpus; and 61.7% of* 1*-grams in the non-ratioed corpus are unique to that corpus.*

"VOTE") tend to appear more often than in ratioed tweets. The imbalance of word counts in the ratioed and non-ratioed tweets is roughly proportional to the tweet imbalance with 22.9% of words occurring in the ratioed tweets. Of all unique words, nearly twice as many occur once or more in the non-ratioed tweets compared to the ratioed tweets.

2.5 Discussion

We have considered how both the volume and kind of user activities in response to US presidents on Twitter varies over multiple time scales. We found the Trump account to have greater variability in the normalized ratio of activities—including a tendency to receive more replies relative to likes and retweets than when compared to the Obama account. Obama's tendency to receive more retweets and likes was only amplified after his departure from public office. We found that in ratioed tweets authored by Trump, words pertaining to fake news, the Mueller inquiry, and the NFL are more common than in non-ratioed tweets. We also showed more general results for response activity profiles, with responses to the two Twitter presidents often stabilizing around the 1-day scale. The Trump account also experiences more activity fluctuations after 1-week than compared to the Obama account, perhaps owing to users more often returning to Trump's comments as political actions unfold.

How the public responds to messages over the course of a politician's career is a fundamental dynamic in politics. As candidates rise to prominence and are elected to office, the reach and importance of their communication changes. Further, the rapid proliferation of digital technology provides a backdrop of constant evolution in polit-

ical communications. Response activity by users on social media provides a valuable set of features for gauging how social networks respond to political messaging. Beyond simple counts, response time series dynamics provide insight into the characteristics of political engagement. This includes determining when activity counts stabilize and perhaps establishing how polarized push-and-pull dynamics play out in response to messages. We speculate these techniques may also be useful for detecting non-human user behavior in response to tweets. When combined with more conventional natural language processing techniques, these "ratiometrics" hold promise in improving our understanding of how specific messaging (e.g., word choice) affects user responses.

This study was not able to control for background changes in the Twitter interface—changes to the manner in which retweets are served could have effects on the response profiles. Further, there are some types of posts—namely "promoted" or "ad" posts—that have different distribution characteristics (i.e. may be viewed by a more targeted, limited audience). We were not able to fully filter these posts and they were included in the analysis.

Twitter is a constantly changing system, and this is especially true for the nature of political discourse on the platform. While we make efforts here to highlight some notable exogenous events, we cannot control for background flux of user-engagement with politics. Simple changes that may be occurring include shifts in the distribution of users from across the political spectrum (i.e., more left- or right-leaning user activity) or more subtle changes in content (e.g., more spam and less in-depth conversation). There is also the potential for modifications to Twitter's timeline algorithms to affect how users engage with content through response activities. Related to this point, we are unsure of algorithmic factors that affect the circulation of inflammatory

content on the platform—it is conceivable that increased views of controversial activities could lead to a snowball effect with heated debates taking place between users in the reply section of a tweet.

With social media playing such a prominent role in today's political process, it is our goal that the methods presented here can contribute to a broader suite of instruments for analyzing political communication in the digital realm. This is the first piece of an effort to systematically evaluate the communications of US presidents on Twitter. Building from our understanding of ratiometrics, the "POTUSometer" is envisioned as sitting alongside instruments such as the Hedonometer [127] (which provides a measurement of collective sentiment on Twitter). The POTUSometer would tap into the 'wisdom of the crowd' (or at least the reality of collective responses) in order to evaluate and better understand how presidents speak and are listened to on social media platforms.

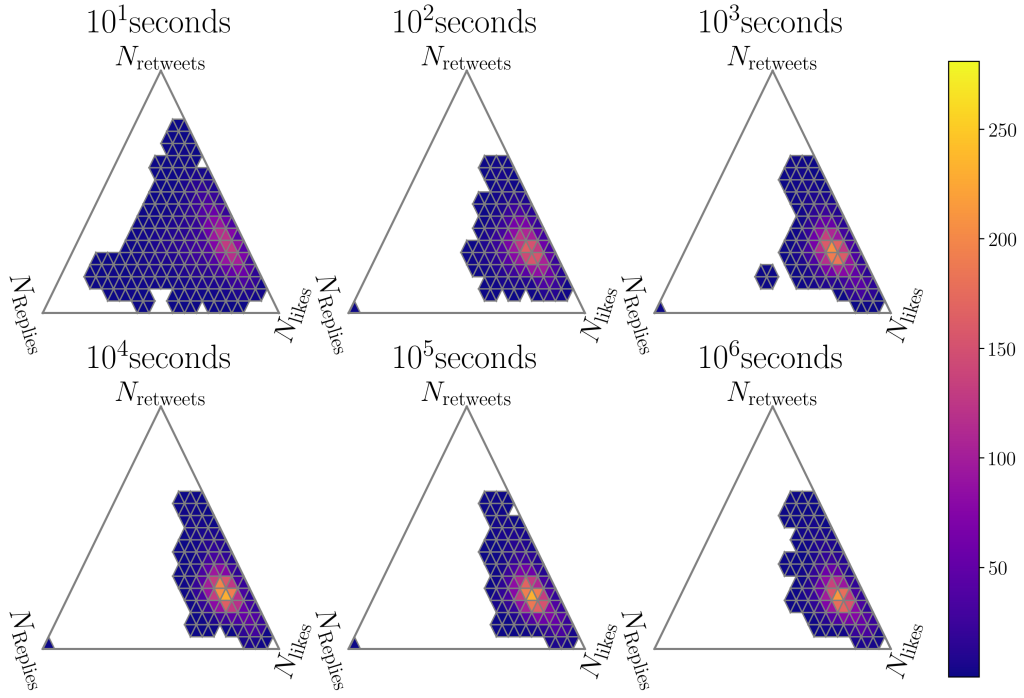
Future research could explore how the sentiment changes with response activities (commented retweets and replies). Similarly, topic modelling could be used to explore which subjects are discussed throughout the response activity time line. A more sophisticated null model for the ternary ratio values (and ratio time series) would be worthwhile to enable anomaly detection. There would be further work in the area of researching attention reinforcement along with analyses of time series stability. With recent advancements in language modelling, text regression on tweet content in order to predict ratio scores may be worthy of investigation. Finally, replicating this work across a broad base of users beyond Obama and Trump would better inform how certain social network attributes effect the above results.

Supporting information

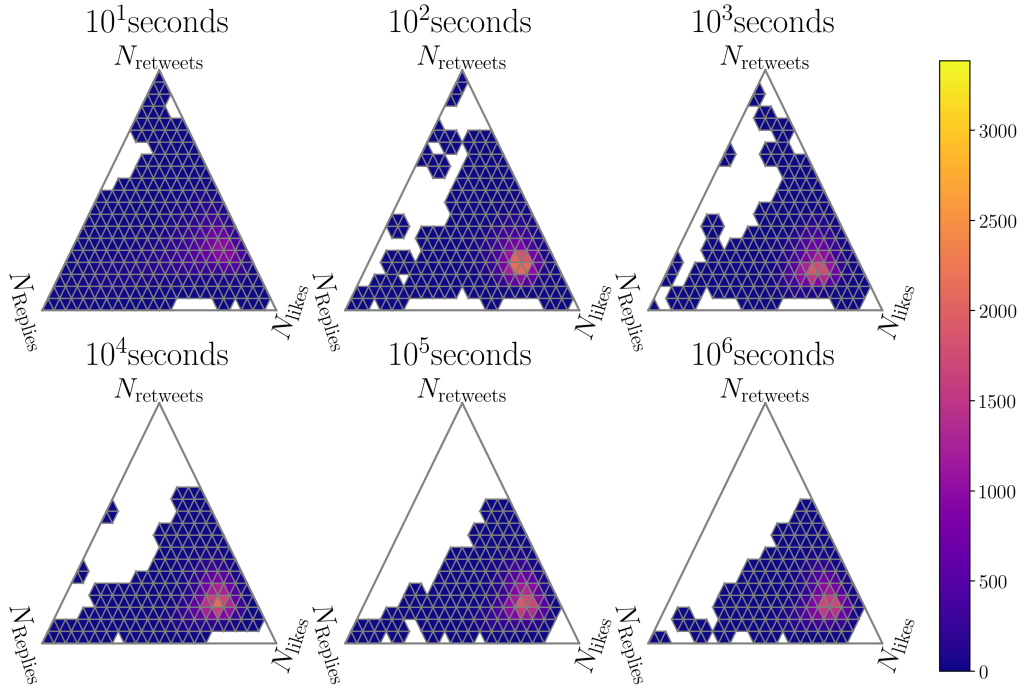
S1 Appendix. Ternary histogram snapshots

Snapshots of ternary histograms are another way of investigating the characteristic time scale described in Sec. 2.4.4 and presented in Fig. 2.6. Here, we calculate the ternary ratios at logarithmically spaced intervals after a tweet is authored (this is in contrast to the stabilized or 'final' ratio values in Fig. 2.3).

S1 Fig. and S2 Fig. show histograms of the ternary ratio values for Obama and Trump tweets at logarithmically spaced time intervals after a tweet is authored. For both presidents there is a greater spread in ternary ratio values immediately after a tweet is authored, with ratios tending to stabilize in the hours and days after a tweet is released. For Trump's tweets, the tendency for ternary ratio values to have a greater reply component is seen throughout the progression of the snapshots. Compared to Obama, Trump tends to have greater spread in the distribution of ternary ratios for each time interval. These results are consistent with our understanding of activity stabilization occurring roughly 1 day after a tweet is authored (as shown in Fig. 2.7).



S1 Fig. Snapshots of ternary activity ratio values for tweets in response to the Obama account. Observations are recorded at logarithmically spaced intervals after the release of the original tweet. Included here are 3,015 tweets from the period of time after Trump's declaration of candidacy on June 16, 2015. Whereas Fig. 2.3 shows a final ratio value for tweets over distinct political periods, here we show how ratios unfold over the life of each tweet. This is serves as a snapshot of the ternary ratio time series presented in Fig. 2.4. Obama's tweets tend to receive a greater volume of likes and retweets relative to replies, and this ratio is often maintained throughout the response timeline.



S2 Fig. Snapshots of ternary activity ratios for tweets in response to the Trump account. Observations are recorded at logarithmically spaced intervals after the release of the original tweet. Included here are 16,708 tweets from the period of time after Trump's declaration of candidacy on June 16, 2015. Whereas Fig. 2.3 shows a final ratio value for tweets over distinct political periods, here we show how ratios unfold over the life of each tweet. This is serves as a snapshot of the ternary ratio time series presented in Fig. 2.4. We can see the greater variation in ternary ratio values compared to the snapshots for Obama (S1 Fig.). There is also a greater tendency for tweets to have higher reply counts when compared to the Obama snapshots.

S3 Fig. Distribution of inverse-inflection points, the local minima of instantaneous retweet volume,

`argrelmin` $\frac{d}{dt}N_{\text{retweets}}$, where $\frac{d^2}{dt^2}N_{\text{retweets}}(t_{i-1}) < 0$ and $\frac{d^2}{dt^2}N_{\text{retweets}}(t_i) > 0$. Here we switch the criteria from Fig. 2.6 in order to show how our choice affects the results for the characteristic time scale. For the results in this figure we show points where the second derivative of the retweets activity time series goes from negative to positive. **A** and **E**: Example cumulative retweet time series and inverse-inflection points (solid triangles) for Obama and Trump tweets. Histograms show the distribution of inverse-inflection points across all tweets binned by time periods before (**B** and **F**), during (**C** and **G**), and after (**D** and **H**) the 2016 US presidential election campaign. The January 1st 2009 to June 15th 2015 period for Trump (**F**) contains inflection point counts that are largely reflective of low initial activity and high(er) late activity (months or years later) leading to unusually high values for seconds to first inflection point ($> 10^8$ seconds). For Obama's and Trump's time in office, tweets experience inflection points around 1-minute and 1-day after the tweet is authored—indicating characteristic time-scales of activity waning. Direct links for Obama tweet (**A**): `https://twitter.com/BarackObama/status/693571153336496128` and Trump tweet (**E**): `https://twitter.com/realDonaldTrump/status/1090729920760893441`. Screenshots were collected on May 28, 2020.

S4 Fig. Obama's ratioed tweets from the study period. Here we show the instances where we observed Obama's tweets garnering more replies than retweets. For the tweets in panels (B, D and G) the response activities later changed to reflect fewer replies than retweets. This behavior is consistent with what we have seen in other cases, with activity shifts either owing to authors deleting the activity or responding-accounts being deleted. Here it seems Obama's tweets are a combination

of controversial (the Iran nuclear deal) and engagement-seeking (soliciting entries in a contest to meet the Obama).

Direct links to tweets for panel **(A)**: https://twitter.com/BarackObama/status/733055491664797696,

panel **(B)**: https://twitter.com/BarackObama/status/639553864006430721,

panel **(C)**: https://twitter.com/BarackObama/status/794926969829920768,

panel **(D)**: https://twitter.com/BarackObama/status/666396723191808000,

panel **(E)**: https://twitter.com/BarackObama/status/652914197454561281,

panel **(F)**: https://twitter.com/BarackObama/status/732691885139984384,

and

panel **(G)**: https://twitter.com/BarackObama/status/715194440487309312.

Tweet screenshots were collected on November 15, 2020.

S2 Appendix. Brief note on bots

In the main body, we discuss the likely influence of bots in the response activity time series. While we do not address this topic in-depth at any point in the present work, we reviewed accounts that retweet to reply to one of two Trump and two Obama tweets in order to gain a preliminary understanding of the volume of bot activity.

Bot detection is a challenging and open-ended task. We use Botometer [128] to score the likelihood that accounts are bots. We randomly select 200 unique users who retweeted highly-liked Trump and Obama tweets. We also randomly select 200 unique users who replied to the same tweets, presenting these results separately. We then run these users through Botometer and report the conditional probability ("cap") that users with a similar score are bots.

For the 200 retweets of Obama tweets, 2.3% had conditional probability scores greater than or equal to 0.8 (cap $\geq$ 0.8). For the 200 retweets of Trump tweets we found that 9.5% of tweets had cap values greater than or equal to 0.8.

For the 200 replies, 16.9% and 8.0% of the Obama and Trump replies, respectively, had conditional probability scores greater than or equal to 0.8.

These results are initial estimates of the overall prevalence of bots in the response activity data. It provides a sense for the order of magnitude of bot activity that may truly be occurring. There is ample room for future research on a larger sample and with more developed methods.

The two Trump tweets examined: https://twitter.com/realDonaldTrump/status/1157345692517634049 and https://twitter.com/realDonaldTrump/status/881503147168071680.

The two Obama tweets examined: https://twitter.com/BarackObama/status/896523232098078720 and https://twitter.com/BarackObama/status/1221552460768202756.

Chapter 3

Mitigating bias in electronic health records

3.1 Abstract

Medical systems in general, and patient treatment decisions and outcomes in particular, can be affected by bias based on gender and other demographic elements. As language models are increasingly applied to medicine, there is a growing interest in building algorithmic fairness into processes impacting patient care. Much of the work addressing this question has focused on biases encoded in language models—statistical estimates of the relationships between concepts derived from distant reading of corpora. Building on this work, we investigate how differences in gender-specific word frequency distributions and language models interact with regards to bias. We identify and remove gendered language from two clinical-note datasets and describe a new debiasing procedure using BERT-based gender classifiers. We show minimal degradation in health condition classification tasks for low- to medium-levels of dataset bias

removal via data augmentation. Finally, we compare the bias semantically encoded in the language models with the bias empirically observed in health records. This work outlines an interpretable approach for using data augmentation to identify and reduce biases in natural language processing pipelines.

3.2 Introduction

Efficiently and accurately encoding patient information into medical records is a critical activity in healthcare. Electronic health records (EHRs) document symptoms, treatments, and other relevant histories—providing a consistent reference through disease progression, provider churn, and the passage of time. Free-form text fields, the unstructured natural language components of a health record, can be incredibly rich sources of patient information. With the proliferation of EHRs, these text fields have also been an increasingly valuable source of data for researchers conducting large-scale observational studies.

The promise of EHR data does not come without apprehension however, as the process of generating and analyzing text data is open to the influence of conscious and unconscious human bias. For example, health care providers entering information may have implicit or explicit demographic biases that ultimately become encoded in EHRs. Furthermore, language models that are often used to analyze clinical texts can encode broader societal biases [49]. As patient data and advanced language models increasingly come into contact, it is important to understand how existing biases may be perpetuated in modern day healthcare algorithms.

In the healthcare context, many types of bias are worth considering. Race, gender, and socioeconomic status, among other attributes, all have the potential to introduce bias into the study and treatment of medical conditions. Bias may manifest in how patients are viewed, treated, and—most relevant here—documented. Due to ethical and legal considerations, as well as pragmatic constraints on data availability, we have focused the current research on gender bias.

There are many sources of algorithmic bias along with multiple definitions of fairness in machine learning [38]. Bias in the data used for training algorithms can stem from imbalances in target classes, how specific features are measured, and historical forces leading certain classes to have longstanding, societal misrepresentation. Definitions of fairness include demographic parity, counterfactual fairness [39], and fairness through unawareness (FTU) [40].

In the current work, we use a more general measure that we refer to as *potential bias* in order to gauge the impact of our data augmentation technique. Potential bias is an assessment of bias under a sort of worst-case scenario, and provides a generalized measure independent of specific bias definitions.

With our methods, we seek to provide human-interpretable insights on potential bias in the case of binary class data. Further, using the same measurement, we experiment with the application of a FTU-like data augmentation process—although the concept of FTU does not neatly translate to unstructured text data (owing to the challenge of neatly and meaningfully redacting linguistic features in textual data). Combined, these methods can identify fundamental bias in language usage and the potential bias resulting from the application of a given machine learning model.

We refer to two classes of algorithmic-bias evaluation: a) intrinsic evaluation for exploring semantic relationships within an embedding space, and b) extrinsic evaluation for determining downstream performance differences on extrinsic tasks (e.g., classification) [129].

In medical context there are gender specific elements that can influence treatment and care. However, in this same context there might also be uses of gender when it is not relevant. In this manuscript we aim to understand the latter through analysis of clinical notes.

There is growing interest in interpretable machine learning (IML) [130]. In the context of deep language models this can involve interrogating the functionality of specific model layers (e.g., BERTology [42]), or investigating the impact of perturbations in data on outputs. This latter approach ties into the work outlined in this manuscript.

Our use of the term 'interpretable' here mostly refers to a more general case where a given result can be interpreted by a human reviewer. For instance, our divergence-based measures highlight gendered terms in plain English with a clearly explained ranking methodology. While this conceptualization is complementary to IML, it does not necessarily fit cleanly within the field—we will mention explicitly when referring to an IML concept.

3.2.1 Prior work

Gender bias in the field of medicine is a topic that must be viewed with nuance in light of the strong interaction between biological sex and health conditions. Medicine

and gender bias interact in many ways—some of which are expected and desirable whereas others may have uncertain or negative impacts in patient outcomes.

Research has reported differences in the care and outcomes received by male and female patients for the same conditions. For example, given the same severity symptoms, men have higher treatment rates for conditions such as coronary artery disease, irritable bowel syndrome, and neck pain [131]. Women have higher treatment-adjusted excess mortality than men when receiving care for heart attacks [132]. Female patients treated by male physicians have higher mortality rates than when treated by female physicians—while male patients have similar mortality regardless of provider gender [133].

The rate of care-seeking behavior in men has been shown to be lower than women and has the potential to significantly affect health outcomes [134]. Some work has shown female providers have higher confidence in the truthfulness of female patients and resulting diagnoses when compared to male providers [135]. The concordance of patient and provider gender is also positively associated with rates of cancer screening [136].

Beyond gender, the mortality rate of black infants has been found to be lower when cared for by black physicians rather than their white counterparts [137]. Race and care-seeking behavior have also been shown to interact, with black patients more often seeking cardiovascular care from black providers than non-black providers [138]. It is important to note historical mistreatment and inequitable access when discussing racial disparities in health outcomes—for instance, the unethical Tuskegee Syphilis Study was found to lead to a 1.5-year decline in black male life expectancy through

increased mistrust in the medical field after the exploitation of its participants was made public [139].

The gender of the healthcare practitioner can also impact EHR note characteristics that are subsequently quantified through language analysis tools. The writings of male and female medical students have been shown to have differences, with female students expressing more emotion and male students using less space [140]. More generally, some work has shown syntactic parsers generalize well for men and women when trained on data generated by women whereas training the tools on data from men leads to poor performance for texts written by women [141].

The ubiquity of text data along with advances in natural language processing (NLP) have led to a proliferation of text analysis in the medical realm. Researchers have used social media platforms for epidemiological research [142–144]—raising a separate set of ethical concerns [145]. NLP tools have been used to generate hypotheses for biomedical research [146], detect adverse drug reactions from social media [147], classes and expand the known lexicon around medical topics [148, 149]. There are numerous applications of text analysis in medicine beyond patient health records. While this manuscript does not directly address tasks outside of clinical notes, it is our hope that the research could be applied to other areas. It is because our methods are interpretable and based on gaining an empirical view of bias that we feel they could be a first resource in understanding bias beyond our example cases of gender in clinical texts.

Our work leverages computational representations of statistically derived relationships between concepts, commonly known as *word embedding* models [150]. These real-valued vector representations of words facilitate comparative analyses of text

data with machine learning methods. The generation of these vectors depends on the distributional hypothesis, which states that similar words are more likely to appear together within a given context. Ideally, word embeddings map semantically similar words to similar regions in the vector space—or 'semantic space' in this case. The choice of training dataset heavily impacts the qualities of the language model and resulting word embeddings. For instance, general purpose language models are often trained on Wikipedia and the Common Crawl collection of web pages (e.g., BERT [24], RoBERTa [151]). Training language models on text from specific domains often improves performance on tasks in those domains (see below). More recent, state-of-the-art word embeddings (e.g., ELMo [152], BERT [24], GPT-2 [25]) are generally 'contextual', where the vector representation of a word from the trained model is dependent on the context around the word. Older word embeddings, such as GloVe [7] and word2vec [6, 13], are 'static', where the output from the trained model is only dependent on the word of interest—with context still being central to the task of training the model.

As medical text data are made increasingly accessible through EHRs, there has been a growing focus on developing word embeddings tailored for the medical domain. The practice of publicly releasing pre-trained, domain-specific word embeddings is common across domains, and it can be especially helpful in medical contexts described using specialized vocabulary (and even manner of writing). SciBERT is trained on a random sample of over one million biomedical and computer science papers [153]. BioBERT similarly is trained on papers from PubMed abstracts and articles [27]. There are also pre-trained embeddings focused on tasks involving clinical notes. Clinical BERT [28, 154] is trained on clinical notes from the MIMIC-III

dataset [155]. A similar approach was applied with the XLNet architecture, resulting in clinical XLNet [156]. These pre-trained embeddings perform better on domain-specific tasks related to the training data and procedure.

The undesirable bias present in word embeddings has attracted growing attention in recent years. Bolukbasi *et al.* present evidence of gender bias in word2vec embeddings, along with proposing a method for removing bias from gender-neutral terms [46]. Contextual word embeddings (e.g., BERT) show gender-biases [157] that can have effects on downstream tasks, although these biases may present differently than those in static embeddings [158,159]. Vig *et al.* investigate which model components (attention heads) are responsible for gender bias in transformer-based language models (GPT-2) [160]. A simple way to mitigate gender bias in word embeddings is to 'swap' gendered terms in training data when generating word embeddings [47]. Beutel *et al.* [161] develop an adversarial system for debiasing language models—in the process, relating the distribution of training data to its effects on properties of fairness in the adversarial system. Simple masking of names and pronouns may reduce bias and improve classification performance for certain language classification tasks [162]. Swapping names has been shown to be an effective data augmentation technique for decreasing gender bias in pronoun resolution tasks [163]. Simple scrubbing of names and pronouns has been used to reduce gender-biases in biographies [164]. Zhang *et al.* examine the gender and racial biases present in Clinical BERT, concluding that after fine-tuning "[the] baseline clinical BERT model becomes more confident in the gender of the note, and may have captured relationships between gender and medical conditions which exceed biological associations." [49] Some of these techniques for

bias detection and mitigation have been critiqued as merely capturing over-simplified dimensions of bias—with proper debiasing requiring more holistic evaluation [48].

Data augmentation has been used to improve classification performance and privacy of text data. Simple methods include random swapping of words, random deletion, and random insertion [165]. More computationally expensive methods may involve using language models to generate contextually accurate synonyms [166], or even running text through multiple rounds of machine translation (e.g., English text to French and back again) [167]. De-identification is perhaps the most common data augmentation task for clinical text. Methods may range from simple dictionary look-ups [168] to more advanced neural network approaches [169]. De-identification approaches may be too aggressive and limit the utility of the resulting data while also offering no formal privacy guarantee. The field of differential privacy [170] offers principled methods for adding noise to data, and some recent work has explored applying these principles to text data augmentation [171]. Applying data-augmentation techniques to pipelines that use contextual word-embeddings presents some additional uncertainty given on-going nature of research working on establishing what these trained embeddings actually represent and how they use contextual clues (e.g., the impact of word order on downstream tasks [172]).

In the present study, we explore the intersection of the bias that stems from language choices made by healthcare providers and the bias encoded in word embeddings commonly used in the analysis of clinical text. We present interpretable methods for detecting and reducing bias present in text data with binary classes. Part of this work is investigating how orthogonal text relating to gender bias is to text related to clinically-relevant information. While we focus on gender bias in health records,

this framework could be applied to other domains and other types of bias as well. In Sec. 3.3, we describe our data and methods for evaluating bias. In Sec. 3.4, we present our results contrasting empirically observed bias in our sample data with bias encoded in word embeddings. Finally, in Sec. 3.5, we discus the implications of our work and potential avenues for future research.

Our main contributions include the following:

- We demonstrate a model-agnostic technique for identifying biases in language usage for clinical notes corresponding to female and male patients. We provide examples of words and phrases highlighted by our method for two clinical note datasets. This methodology could readily be applied to other demographic attributes of patients as well.

- Continuing with the bias identification technique, we contrast the results from the model-agnostic bias detection with results from evaluating bias within a word-embedding space. We find that our model-agnostic method highlights domain- and dataset-specific terms, leading to more effective bias identification than when compared with results derived from language models.

- We develop a data augmentation procedure to remove biased terms and present results demonstrating that this procedure has minimal effect on clinically relevant tasks. Our experiments show that removing words corresponding to 10% of the language-distribution divergence has little effect on condition classification performance while largely reducing the gender signal in clinical notes. Further, the augmentation procedure can be applied to a high volume of terms (for terms corresponding to up to 80% of total language-distribution divergence)

with minimal degradation in performance for clinically relevant tasks. More broadly, our results demonstrate that transformers-based language models can be robust to high levels of data augmentation—as indicated by retention of relative performance on downstream tasks.

Taken together, these contributions provide methods for bias identification that are readily interpreted by patients, providers, and healthcare informaticians. The bias measures are model-agnostic and dataset specific and can be applied upstream of any machine learning pipeline. This manuscript directly supports the blossoming field of ethical artificial intelligence in healthcare and beyond. Our methods could be helpful for evaluating the impact of demographic signals—beyond gender—present in text when developing machine learning models and workflows in healthcare.

3.3 Methods

Here we outline our methods for identifying and removing gendered language and evaluating the impact of this data augmentation process. We also provide brief descriptions of the datasets for our case study. The bias evaluation techniques fall into two main categories. First, we make intrinsic evaluations of the language by looking at bias within word-embedding spaces and empirical word-frequency distributions of the datasets. The set of methods presented here enable the identification of biased language usage, a data augmentation approach for removing this language, and an example benchmark for evaluating the performance impacts on two biomedical datasets. Second, there are extrinsic evaluation tasks focused on comparing the performance of classifiers as we vary the level of data augmentation. For our dataset, this process

involves testing health-condition and gender classifiers on augmented data. The extrinsic evaluation provides a measurement of potential bias and is meant to be similar to some real-world tasks that may utilize similar classifiers.

3.3.1 Bias measurements

We define three different bias evaluation frameworks for our study. The first is a measure of empirical bias between language distributions for two classes observed in a given dataset. The second is a measure of intrinsic bias present in a word embedding space (as explored with a given corpus). Finally, the third measure addresses the *potential* extrinsic algorithmic-bias present in a machine learning pipeline.

For our evaluations of intrinsic language bias in empirical data we use a divergence metric (details in Sec. 3.3.2) which is calculated between the language distributions of male and female patients. We use a straightforward notion of bias that rates n-grams as more biased when their divergence contribution is higher. In this case we are detecting data bias, which may in itself have multiple sources such as measurement or sampling biases.

We evaluate intrinsic language bias in embedding spaces by calculating similarity scores between gendered word-embedding clusters and n-grams appearing in male and female patient notes (details in Sec. 3.3.4). Here we are focused on a measure of algorithmic bias as expressed via language from a specific dataset that is encoded with a given language model.

Finally, we evaluate extrinsic bias to test the effects of our data augmentation procedure. In this case we diverge from the largely established definitions of bias by using an evaluation framework that does not claim to detect explicit forms of bias for

protected classes. Instead, we reframe our extrinsic measure as *potential bias* (PB), which we define generally as the capacity for a classifier to predict protected classes. Our PB measure does not equate directly to real-world biases, but we argue that it has utility in establishing a generalizeable indication of the potential for bias that is task-independent. Stated another way, it is a measure of the signal present in a data set and the capacity for a given machine learning algorithm to utilize this signal for biased predictions. In our case, we present PB as measured by the performance of a binary classifier trained to predict patient gender from text documents.

We recognize concerns raised by Blodgett *et al.* [173] and others relating to the imprecise definitions of bias in the field of algorithmic fairness. Indeed, it is often important to motivate a given case of bias by establishing its potential harms. In our case, we are limiting what we present to precursors to bias, and thus do not claim to be making robust assessments of real-world bias and subsequent impact.

3.3.2 Rank-divergence and trimming

We parse clinical notes into n-grams—sequences of space delimited strings such as words or sentence fragments—and generate corresponding frequency distributions. To quantify the bias of specific n-grams we compare their frequency of usage in text data corresponding to each of our two classes. The same procedure is extended as a data augmentation technique intended to remove biased language in a targeted and principled manner.

More specifically, in our case study of gendered language in clinical notes we quantify the "genderedness" of n-grams by comparing their frequency of usage in notes for male and female patient populations. For the task of comparing frequency

of n-gram usage we use rank-turbulence divergence (RTD), as defined by Dodds *et al.* [3]. The rank-turbulence divergence between two sets, Ω_1 and Ω_2, is calculated as follows,

$$\begin{aligned} D_\alpha^R(\Omega_1||\Omega_2) &= \sum \delta D_{\alpha,\tau}^R \\ &= \frac{\alpha+1}{\alpha} \sum_\tau \left| \frac{1}{r_{\tau,1}^\alpha} - \frac{1}{r_{\tau,2}^\alpha} \right|^{1/(\alpha+1)} , \end{aligned} \tag{3.1}$$

where $r_{\tau,s}$ is the rank of element τ (n-grams, in our case) in system s and α is a tunable parameter that affects the starting and ending ranks. While other techniques could be used to compare the two n-gram frequency distributions, we found RTD to be robust to differences in overall volume of n-grams for each patient population. For example, Fig. 3.1 shows the RTD between 1-grams from clinical notes corresponding to female and male patients.

A brief note on notation: τ always represents a unique element, or n-gram in our case. In certain contexts τ may be an integer value that ultimately maps back to the element's string representation. This integer conversion is to allow for clean indexing—in these cases τ can be converted back to a string representation of the element with the array of element strings W_τ.

We use the individual rank-turbulence divergence contribution, $\delta D_{\alpha,\tau}^R$, of each 1-gram to the gendered divergence, $D_\alpha^R(\Omega_{\text{female}}||\Omega_{\text{male}})$, to select which terms to remove from the clinical notes. First, we sort the 1-grams based on their rank-turbulence divergence contribution. Next, we calculate the cumulative proportion of the overall rank-turbulence divergence, RC_τ, that is accounted for as we iterate through the sorted 1-gram list from words with the highest contribution to the least contribution (in this case, terms like "she" and "gentleman" will tend to have a greater contribu-

tion). Finally, we set logarithmically spaced thresholds of cumulative rank-divergence values to select which 1-grams to trim. The method allows us to select sets of 1-grams that contribute the most to the rank-divergence values (measured as divergence per 1-gram). Fig. 3.2 provides a graphical overview of this procedure.

Using this selection criteria, we are able to remove the least number of 1-grams per a given amount of rank-turbulence divergence removed from the clinical notes. The number of unique 1-grams removed per cumulative amount of rank-turbulence divergence grows super linearly as seen in Fig. 3.8I. This results in relatively stable distributions of document lengths for lower trim values (10–30%), although at higher trim values the procedure drastically shrinks the size of many documents (Fig. 3.8A-H).

To implement this trimming procedure, we use regular expressions to replace the 1-grams we have identified for removal with a space character. We found that using 1-grams as the basis for our trimming procedure is both effective and straightforward to implement. Generally, if higher order n-grams (e.g., 2-grams) are determined to be biased, the constituent 1-grams are also detected by the RTD metric. Our string removal procedure is applied to the overall corpus of data, upstream of any train-test dataset generation for specific classification tasks.

Other potential string replacement strategies include redaction with a generic token or randomly swapping n-grams that appear within the same category across the corpus [171]. The RTD method we use could also be adapted for use with these and other replacement strategies. We chose string removal because of its simplicity and prioritization of the de-biasing task over preserving semantic structure (i.e., it presents an extreme case of data augmentation). The pipeline's performance on

downstream tasks provides some indication of the semantic information retained, and as we show in Sec. 3.4 it is possible retain meaningful signals while pursuing relatively aggressive string removal.

Algorithm 1 RTD trimming procedure

Input: Documents $D_i \quad , i = 1, ..., N$
Input: 1-gram rank dists. for each class $\Omega_\psi \quad , \psi = 1, 2$
Output: Trimmed text data $C_i^{(k)} \quad , i = 1, ..., N \, ; k \in (0, 1]$

```
δD^R_{α,τ}, W_τ ← RTD_calc(Ω_1, Ω_2, α)   , τ = 1, ..., M   ▷ δD^R_{α,τ} is RTD contribution for ngram W_τ, both sorted by RTD contribution
RC_τ ← cumsum(δD^R_{α,τ})
for k = .1, .2, ..., .9 do
    r ← max(where(RC <= k))                                  ▷ index up to bin max b
    S ← W_{0:r}
    for i = 1, ..., N do
        C_i^(k) ← strip(D_i, S)                              ▷ remove 1−grams from doc.
    end for
end for
```

3.3.3 Language models

Large language models are increasingly common in many NLP tasks, and we feel it is important to present our results in the context of a pipeline that utilizes these models. Furthermore, language models have the potential to encode bias, and we found it necessary to contrast our empirical PB detection methods with bias metrics calculated on general purpose and domain-adapted language models.

We use pre-trained BERT-base [24] and Clinical BERT [154] word embeddings. BERT provides a contextual word embedding trained on "general" language whereas Clinical BERT builds on these embeddings by utilizing transfer learning to improve

performance on scientific and clinical texts. All models were implemented in PyTorch using the Transformers library [174].

For tasks such as nearest-neighbor classification and gender-similarity scoring, we use the off-the-shelf weights for BERT and Clinical BERT (see Fig. 3.3 for an example of n2c2 embedding space). These models were then fine-tuned on the gender and health-condition classification tasks.

In cases where we fine-tuned the model, we added a final linear layer to the network. All classification tasks were binary with a categorical cross-entropy loss function. All models were run with a maximum sequence length of 512, batch size of 4, and gradient accumulation steps set to 12. We considered various methods for handling documents longer than the maximum sequence length (see *Variable length note embedding* in SI), but ultimately the performance gains did not merit further use.

We also run a nearest-neighbor classifier on the document embeddings produced by the off-the-shelf BERT-base and Clinical BERT models. This is intended to be a point of comparison when evaluating the potential bias present within the embedding space, as indicated by performance on extrinsic classification tasks.

In addition to the BERT-based language models, we used a simple term frequency-inverse document frequency (TFIDF) [175] based classification model as a point of comparison. For this model, we fit a TFIDF vectorizer to our training data and use logistic regression for binary classification.

For classification performance metrics we report the Matthews correlation coefficient (MCC) and receiver operating characteristic (ROC) curves. We primarily use

MCC values for ease of presentation and because of the balanced nature of the measurement even in the face of class imbalances [176]. MCC is calculated as follows,

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{3.2}$$

Where TP, TN, FP,and FN are counts of true positives, true negatives, false positives, and false negatives, respectively. MCC ranges between -1 and 1 with 1 indicating the best performance and 0 indicating performance no better than random guessing. The measure is often described as the correlation between observed and predicted labels.

3.3.4 Gender distances in word embeddings

Using a pre-trained BERT model, we embed all the 1-grams present in the clinical note datasets. For this task, we retain the full length vector for each 1-gram, taking the average in cases where additional tokens are created by the tokenizer. The results of this process are 1x768 vectors for each n-gram. We also calculate the average embedding for a collection of terms manually selected to constitute 'gender clusters'. From these gender clusters we calculate the cosine similarly to each of the embeddings for n-grams in the Zipf distribution.

Using measures such as cosine similarity with BERT raises some concerns, especially when looking at absolute values. BERT was not trained on the task of calculating sentence or document similarities. With BERT-base, all dimensions are weighted equally, which when applying cosine similarity can result in somewhat arbitrary absolute values. As a workaround, we believe that using the ranked value of

word and document embeddings can produce more meaningful results (if we do not wish to fine-tune BERT on the task of sentence similarity). We use both absolute values and ranks of cosine similarity when investigating bias in BERT-based language models—finding the absolute values of cosine similarity to be meaningful in our relatively coarse-grained analysis. Further, taking the difference in cosine similarities for each gendered cluster addresses some of the drawbacks of examining cosine similarity values in pre-trained models.

Generating word or phrase embeddings from contextual language models raises some challenges in terms of calculating accurate embedding values. In many cases, the word-embedding for a given 1-gram—produced by the final layer of a model such as BERT—can vary significantly depending on context [177]. Some researchers have proposed converting contextual embeddings to static embeddings to address this challenge [178]. Others have presented methods for creating template sentences and comparing the relative probability of masked tokens for target terms [158]. After experimenting with the template approach, we determined that the resulting embeddings were not different enough to merit switching away from the simple isolated 1-gram embeddings.

3.3.5 Rank-turbulence divergence for embeddings and documents

We use rank-turbulence divergence in order to compare the bias encoded in word-embeddings and empirical data. For word embeddings, we need to devise a metric for bias—here we use cosine similarity between biased-clusters and candidate n-grams.

The bias in the empirical data is evaluated using RTD for word-frequency distributions corresponding to two labeled classes.

In terms of the clinical text data, for the word embeddings we use cosine similarity scores to evaluate bias relative to known gendered n-grams. For the clinical note datasets (text from documents with gender labels), we use rank-turbulence divergence calculated between the male and female patient populations.

To evaluate bias in the embedding space, we rely on similarity scores relative to known gendered language. First, we create two gendered clusters of 1-grams—these clusters represent words that are manually determined to have inherent connotations relating to female and male genders. Next, we calculate the cosine similarity between the word embeddings for all 1-grams appearing in the empirical data and the average vector for each of the two gendered clusters. Finally, we rank each 1-gram based on the distribution of cosine similarity scores for the male and female clusters.

For the empirical data, we calculate the RTD for 1-grams appearing in the clinical note data sets. The RTD value provides an indication of the bias—as indicated by differences in specific term frequency—present in the clinical notes.

Combined, these steps provide ranks for each 1-gram in terms of how much it differentiates the male and female clinical notes. Here again we can use the highly flexible rank-turbulence divergence measure to identify where there is 'disagreement' between the ranks returned by evaluating the embedding space and ranks from the empirical distribution. This is a divergence-of-divergence measure, using the iterative application of rank-turbulence divergence to compare two different measures of rank. Going forward, we refer to this measure as RTD^2. RTD^2 provides an indication of which n-grams are likely to be reported as less gendered in either the embedding space

or in the empirical evaluation of the documents. For our purposes, RTD^2 is especially useful for highlighting n-grams that embedding-based debiasing techniques may rank as minimally biased, despite the emperical distribution suggesting otherwise.

3.3.6 Data

We use two open source datasets for our experiments: the n2c2 (formerly i2b2) 2014 deidentification challenge [179] and the MIMIC-III critical care database [155]. The n2c2 data comprises around 1300 documents with no gender or health condition coding (we generate our own labels for the former). MIMIC-III is a collection of diagnoses, procedures, interventions, and doctors notes for 46,520 patients that passed through an intensive care unit. There are 26,121 males and 20,399 females in the dataset, with over 2 million individual documents. MIMIC-III includes coding of health conditions with International Classification of Diseases (ICD-9) codes, as well as patient sex.

For MIMIC-III, we focus our health-condition classification experiments on records corresponding to patients with at least 1 of the top 10 most prevalent ICD-9 codes. We restrict our sample population to those patients with at least one of the 10 most common health conditions—randomly drawing negative samples from this subset for each condition classification experiment. Rates of coincidence vary between 0.65 and 0.13 (Fig. 3.9). All but one of the top 10 health conditions have more male than female patients (Table 3.1). As a point of reference, we also present summary results for records corresponding to patients with ICD-9 codes that appear at least 1000 times in the MIMIC-III data (Table 3.8).

ICD Description.	Sex Count N_{f}	N_{m}	$N_{\mathrm{f}}/N_{\mathrm{total}}$	$N_{\mathrm{m}}/N_{\mathrm{total}}$
Acute kidney failure	3941	5178	0.43	0.57
Acute respiratory failure	3473	4024	0.46	0.54
Atrial fibrillation	5512	7379	0.43	0.57
Congestive heart failure	6106	7005	0.47	0.53
Coronary atherosclerosis	4322	8107	0.35	0.65
Diabetes mellitus	3902	5156	0.43	0.57
Esophageal reflux	2990	3336	0.47	0.53
Essential hypertension	9370	11333	0.45	0.55
Hyperlipidemia	3537	5153	0.41	0.59
Urinary tract infection	4027	2528	0.61	0.39

Table 3.1: ***Patient sex ratios for the top 10 conditions in MIMIC-III.*** *For most health conditions there is an a imbalance in the gender ratio between male and female patients. This reflects an overall bias in the MIMIC-III dataset witch has more male patients.*

3.3.7 Text pre-processing

Before analyzing or running the data through our models, we apply a simple pre-processing procedure to the text fields of the n2c2 and MIMIC-III data sets. We remove numerical values, ranges, and dates from the text. This is done in an effort to limit confounding factors related to specific values and gender (e.g., higher weights and male populations). We also strip some characters and convert common abbreviations. See Sec. 3.7.1 for information on note selection.

3.4 Results

Here we present the results of applying empirical bias detection and potential bias mitigation methods. Using rank-turbulence divergence (RTD) we rank n-grams based on their contribution to bias between two classes. Next we apply a data augmenta-

tion procedure where we remove 1-grams based on their ranking in the RTD results. The impact of the data augmentation process is measured by tracking classification performance as we apply increasingly aggressive 1-gram trimming to our clinical note datasets. Finally, we compare the bias present in the BERT embedding space with the empirical bias we detect in the case study datasets.

One of our classification tasks is predicting patient gender from EHR notes. We include gender classification as a synthetic test that is meant to directly indicate the gender signal present in the data. Gender classification is an unrealistic task that we would not expect to see in real-world applications, but serves as an extreme case that provides insight on the potential for other classifiers to incorporate gender information (potential bias).

We present results for both the n2c2 and MIMIC-III datasets. The n2c2 dataset provides a smaller dataset with more homogeneous documents and serves as a reference point for the tasks outlined here. MIMIC-III is much larger and its explicit coding of health conditions allows us to bring the extrinsic task of condition classification into our evaluation of bias and data augmentation.

3.4.1 Gender divergence

Interpretability is a key facet of our approach to empirical bias detection. To gain an understanding of biased language usage we start by presenting the ranks of RTD values for individual n-grams in text data corresponding to each of the binary classes. The allotaxonographs we use to present this information (e.g., Fig. 3.1) show both the RTD values and a 2-d rank-rank histogram for n-grams in each class. The rank-rank histogram (Fig. 3.1 left) is useful for evaluating how the word-frequency distributions

(Fig. 3.1 right) are similar or disjoint among the two classes, and in the process visually inspecting the fit of the tunable parameter α, which modulates the impact of lowly-ranked n-grams. See Figs. 3.12, 3.13, 3.14, and 3.15 for additional allotaxonographs, including 2- and 3-grams.

In the case of our gender classes in the medical data, we find the rank distributions to be more similar than disjoint and visually confirm that $\alpha = 1/3$ is an acceptable setting (by examining the relation between contour lines and rank-rank distribution).

More specifically, in our case study gendered language is highlighted by calculating the RTD values for male and female patient notes. We present results from applying our RTD method to 1-grams in the unmodified MIMIC-III dataset in Fig. 3.1. Unsurprisingly, gendered pronouns appear as the greatest contribution to RTD between the two corpora. Further, 1-grams regarding social characteristics such as "husband", "wife", and "daughter" and medically relevant terms relating to sex-specific or sex-biased conditions such as "hysterectomy", "scrotal", and "parathyroid" are also highlighted.

Some of these terms may be obvious to readers—suggesting the effectiveness of this approach to capture intuitive differences. Upon deeper investigation, 1-grams such as "husband" and "wife" often appear in reference to a patient's spouse providing information or social histories. The reasons for "daughter" appearing more commonly in female patient notes are varied, but appear to be related to higher relative rates of daughters providing information for their mothers. However the identification and ranking of other n-grams in terms of gendered-bias requires examination of a given dataset—perhaps indicating unintuitive relationships between terms and gendered language, or potentially indicating overfitting of this approach to specific datasets.

For instance, "parathyroid" likely refers to hypoparathyroidism which is not a sex-specific condition, but rather a sex-biased condition with a ratio of 3.3:1 for female to male diagnoses. Further, men are more likely to present asymptomatically, which may be less likely to be diagnosed in an ICU setting [180].

The application of RTD produces, in a principled fashion, a list of target terms to remove during the debiasing process—automating the selection of biased n-grams and tailoring results to a specific dataset. Using the same RTD results from above, we apply our trimming procedure—augmenting the text by iteratively removing the most biased 1-grams. For instance, in the MIMIC-III data the top 268 1-grams account for 10% of the total RTD identified by method—and these are the first words we trim.

3.4.2 Gender classification

As an extrinsic evaluation of our biased-language removal process, we present performance results for classifiers predicting membership in the two classes that we obscure through the data augmentation process. We posit that the performance of a classifier in this case is an important metric when determining if the data augmentation was successful in removing bias signals and thus potential bias from the classification pipeline. The performance of the classifier is analogous to a real-world application under an extreme case where we are trying to predict the protected class.

We evaluate the performance of a binary gender classifier based on BERT and Clinical BERT language models. As a starting point, we investigate the performance of a basic nearest neighbors classifier running on document embeddings produced with off-the-shelf language models. The classification performance of the nearest-neighbor classifier is far better than random and speaks to the embedding space's local clus-

tering by gendered of words in these datasets, suggesting that gender may be a major component of the words embedded within this representation space. The tendency for BERT, and to a lesser extent Clinical BERT, to encode gender information can be seen in the tSNE visualization of these document embeddings (Figs. 3.3 and 3.17). As seen here and in other results, Clinical BERT exhibits less potential gender-bias according to our metrics. We leave a more in-depth comparison of gender-bias in BERT and Clinical BERT to future work, but it is worth noting that different embeddings appear to have different levels of potential gender-bias. Further, clinical text data may be more or less gender-biased than everyday text.

The performance of the BERT-based nearest neighbor classifier on the gender classification task is notable (Matthews correlation coefficient of 0.69) given the language models were not fine-tuned (Table 3.2). Using Clinical BERT embeddings result in a MCC of 0.44 for the nearest neighbor classifier—with Clinical BERT generally performing slightly worse on gender classification tasks.

As a point of comparison, we attempt a naive approach to removing gender bias through data augmentation that involves trimming a manually selected group of 20 words. When we run our complete BERT classifier, with fine tuning, for 1 epoch we find that the MCC drops from 0.94 to -0.06 when we trim the manually selected words. This patterns holds up for Clinical BERT as well. However, if we extend the training run to 10 epochs, we find that most of the classification performance is recovered. This suggests that although the manually selected terms may have some of the most prominent gender signals, removing them by no means prevents the models from learning other indicators of gender.

	BERT		Clinical BERT	
Model notes	Gendered	No-gend.	Gendered	No-gend.
Nearest neighbor	0.69	*	0.44	*
1 Epoch	0.94	-0.06	0.92	0.00
10 Epochs	*	0.88	*	0.56

Table 3.2: ***Matthews Correlation Coefficient for gender classification task on n2c2 dataset.*** *BERT and Clinical BERT based models were run on the manually generated "no gender" test dataset (common pronouns, etc. have been removed). The nearest neighbor model uses off-the-shelf models to create document embeddings, while the models run for 1 and 10 Epochs were fine tuned.*

On the MIMIC-III dataset we find gender classification to be generally accurate. With no gender trimming applied, MCC values are greater than 0.9 for both BERT and Clinical BERT classifiers. This performance is quickly degraded as we employ our trimming method (Fig. 3.5K). When we remove 1-grams accounting for the first 10% of the RTD, we find a MCC value of approximately 0.2 for the gender classification task. The removal of the initial 10% of rank-divergence contributions has the most impact in terms of classification performance. Further trimming does not reduce the performance as much until 1-grams accounting for nearly 80% of the rank-turbulence divergence are removed. At this point, the classifier is effectively random, with a MCC of approximately 0.

Taken together these results point to a reduction in potential bias through our trimming procedure.

The large drop in performance for gender classification is in contrast to that of most health conditions (Fig. 3.4B). On the health condition classification task most trim values result in negligible drops in classification performance.

3.4.3 Condition Classification

To evaluate the impact of the bias removal process, we track the performance of classification tasks that are not explicitly linked to the two classes we are trying to protect. Under varying levels of data augmentation we train and test multiple classification pipelines and report any degradation in performance. These tasks are meant to be analogous to real world applications in our domain that would require the maintenance of clinically-relevant information from the text—although we make no effort to achieve state-of-the-art results (see Table 3.3 for baseline condition classification performance).

ICD9 Description	**ICD9 Code**	**MCC**
Diabetes mellitus	25000	0.53
Hyperlipidemia	2724	0.46
Essential hypertension	4019	0.41
Coronary atherosclerosis	41401	0.67
Atrial fibrillation	42731	0.53
Congestive heart failure	4280	0.51
Acute respiratory failure	51881	0.43
Esophageal reflux	53081	0.43
Acute kidney failure	5849	0.29
Urinary tract infection	5990	0.23

Table 3.3: ***Clinical BERT performance on top 10 ICD9 codes in the MIMIC-III dataset.***

In the specific context of our case study, we train health-condition classifiers that produce modest performance on the MIMIC-III dataset. This performance is suitable for our purposes of evaluating the degradation in performance on the extrinsic task, relative to our trimming procedure.

In the case of each health condition, we find that relative classification performance is minimally affected by the trimming procedure. For instance, the classifier for atrial fibrillation results in a MCC value of around 0.48 for the male patients (Fig. 3.5C) in the test set when no trimming is applied. When the minimal level of trimming is applied (10% of RTD removed), the MCC for the males is largely unchanged, resulting in a MCC of 0.48. This largely holds true for most of the trimming levels, across the 10 conditions we evaluate in-depth. For 6 out of 10 conditions, we find that words accounting for approximately 80% of the gender RTD need to be removed before there is a noteworthy degradation of classification performance. At the 80% trim level, the gender classification task has a MCC value of approximately 0, while many other conditions maintain some predictive power.

Comparing the relative degradation in performance, we see that the proportion of MCC lost between no- and maximum-trim between 0.05 and 0.4 for most conditions (Fig. 3.4B). The only condition with full loss of predictive power is for urinary tract infections, which one might also speculate to be related to the anatomical differences in presentation of UTIs between biological sexes. Although, this task also proved the most challenging and had the worst starting (no-trim) performance (MCC ≈ 0.2).

The above results suggest that, for the conditions we examined, performance for medically relevant tasks can be preserved while reducing performance on gender classification. There is the chance that the trimming procedure may result in biased preservation of condition classification task performance. To investigate this we present results from a lightweight, TF-IDF based classifier for 123 health conditions. We find that when we trim the top 50% of RTD that classifiers for most conditions are relatively unaffected (Fig. 3.6). For those conditions that do experience shifts in

classification performance, any gender imbalance appears attributable related to the background gender distribution in the dataset.

3.4.4 Gender distance

To connect the empirical data with the language models, we embed n-grams from our case study datasets and evaluate their intrinsic bias within the word-embedding space. These language models have the same model-architectures that we (and many others) use when building NLP pipelines for classification and other tasks. Bias measures based on the word-embedding space are meant to provide some indication of how debiasing techniques that are more language model-centric would operate (and what specific n-grams they may highlight)—keeping with our theme of interpretability while contrasting these two approaches.

In the context of our case study, we connect empirical data with word embeddings by presenting the distributions of cosine similarity scores for 1-grams relative to gendered clusters in the embedding space. Cosine similarity scores are calculated for all 1-grams relative to clusters representing both female and male clusters (defined by 1-grams in Table. 3.4). In our results we use both the maximum cosine similarity value relative to these clusters (i.e., the score calculated against either the female or male cluster) as well as differences in the scores for each 1-gram relative both female and male clusters. Looking at the distributions of maximum cosine similarity scores for 1-grams appearing in both the n2c2 dataset (Fig. 3.16) and the MIMIC-III dataset (Fig. 3.7B) we observed a bimodal distribution of values. In both figures, a cluster with a mean around 0.9 is apparent as well as a cluster with a mean around 0.6. Through manual review of the 1-grams, we find that the cluster around 0.9 is largely

comprised of more common, conversational English words whereas the cluster around 0.6 is largely comprised of medical terms. While there are more unique 1-grams in the cluster of medical terms, the overall volume of word occurrences is far higher for the conversational cluster.

Referencing the cosine similarity clusters against the rank-turbulence divergence scores for the two data sets, we find that a high volume of individual 1-grams that are trimmed are present in the conversational cluster. However, the number of unique terms there are removed for lower trim-values are spread throughout the cosine-similarity gender distribution. For instance, when trimming the first 1% of RTD, we find that terms are selected in both the more conversational cluster and the more technical cluster (Fig. 3.7E), with the former accounting for far more of the total volume of terms removed. The total volume of 1-grams is skewed towards the conversational cluster with terms that have higher gender similarity (Fig. 3.7G). The fact that the terms selected for early stages of trimming appear across the distribution of cosine similarity values illustrates the benefits of our empirical method, which is capable of selecting terms specific to a given dataset without relying on information contained in a language model. The contrast between the RTD selection criteria and the bias present in the language model helps explain why performance on the condition classification task is minimally impacted even when a high volume of 1-grams are removed—with RTD selecting only the most empirically biased terms. Using RTD-trimming there is a middle ground between obscuring gender and barely preserving performance on condition classifications—some of the more nuanced language can be retained using our method.

3.4.5 Comparison of language model and empirical bias

Finally, we identify n-grams that are more biased in either the language model or in the empirical data, using RTD to divert attention away from n-grams that appear to exhibit similar levels of bias in both contexts. Put more specifically, the first application of RTD—on the empirical data and word-embeddings—ranks n-grams that are more male or female biased. The second application, the *divergence-of-divergence* (RTD^2), ranks n-grams in terms of where there is most disagreement between the two bias detection approaches.

For the MIMIC-III dataset, we find RTD^2 highlights sex-specific terms, social information, and medical conditions (Table 3.6). The abbreviations of "f" and "m" for instance are rank 6288 and 244, respectively for RTD bias measures on BERT. Moving to RTD bias measurements in MIMIC-III, "f" and "m" are the 3rd and 7th most biased terms, respectively, appearing in practically every note when describing basic demographic information for patients. The BERT word embedding of the 1-gram "grandmother" has a rank of 4 but a rank of 3571 in the MIMIC-III data—due to the fact that the 1-gram "grandmother" is inherently semantically gendered, but in the context of health records does not necessarily contain meaningful information on patient gender. "Husband" on the other hand does contain meaningful information on the patient gender (at least in the MIMIC-III patient population), with it being rank 4 in terms of its empirical bias—the word embedding suggests it is biased, but less so with a rank of 860.

As a final set of examples, we look at medical conditions. It is worth noting our choice of BERT rather than Clinical BERT most likely results in less effective word embeddings for medical terms. "Cervical" has a rank of 7 in the BERT bias rankings and a rank of 18374 in the empirical bias distribution—most likely owing to the split meanings in a medical context. Conversely, "flomax" has a rank of 10891 for the word embedding bias, while the empirical bias rank is 11—most likely due to the gender imbalance in the incidence of conditions (e.g., kidney stones, chronic prostatitis) that flomax is often prescribed to treat. Similarly, "hypothyroidism" is ranked 12 in MIMIC and 17831 in BERT RTD ranks, with the condition having a known increased prevalence in female-patients.

The high RTD^2 ranks for medical conditions somewhat owe to the fact that we used BERT rather than the medically-adapted Clinical BERT for these results. For these results the choice to use the general purpose BERT rather than Clinical BERT was motivated by illustrating the discrepancies in bias rankings when using the general purpose model (with the added contrast of a shifted domain, as indicated by jargonistic medical conditions). When applying this type of comparison in practice, it will most likely be more beneficial to compare bias ranks with language models that are used in any final pipeline (in this case, Clinical BERT). Additionally, the difficulty of constructing meaningful clusters of gendered terms using technical language limits the utility of the our cosine similarity bias measure in the Clinical BERT embedding space (see Table 3.7). Inspection of the 1-grams with high RTD^2 values for BERT suggests a word of caution when using general purpose word embeddings on more technical datasets, while also illustrating how specific terms that drive bias may differ between different domains. The lesson derived by the case study of applying

BERT to medical texts could be expanded to provide further caution when working in domains that do not have the benefit of fine-tuned models or where model fit may be generally poor for other reasons.

3.5 Concluding remarks

Here we present interpretable methods for detecting and reducing bias in text data. Using clinical notes and gender as a case study, we explore how using our methods to augment data may affect performance on classification tasks, which serve as extrinsic evaluations of the potential-bias removal process. We conclude by contrasting the inherent bias present in language models with the bias we detect in our two example datasets. These results demonstrate that it is possible to obscure gender-features while preserving the signal needed to maintain performance on medically relevant classification tasks.

Our methods start by using a divergence measure to identify empirical data bias present in our EHR datasets. We then assess the intrinsic bias present within the word embedding spaces for general purpose and clinically adapted language models. We introduce the concept of potential bias (PB) and evaluate the reduction of extrinsic PB when we apply our mitigation strategy. PB results are generated by presenting performance on a gender classification task. Finally, we compare the results of assessing empirical data bias and intrinsic embedding space bias by contrasting the rankings of 1-grams produced by each method.

When evaluating the differences in word use frequency in medical documents, certain intuitive results emerge: practitioners use gendered pronouns to describe pa-

tients, they note social and family status, and they encode medical conditions with known gender imbalances. Using our rank-turbulence divergence approach, we are able to evaluate how each of these practices, in aggregate, contribute to a divergence in word-frequency distributions between the male- and female-patient notes. This becomes more useful as we move to identifying language that while not explicitly gendered may still be used in an unbalanced fashion in practice (for instance, non-sex specific conditions that are diagnosed more frequently in one gender). The results from divergence methods are useful for both understanding differences in language usage and as a debiasing technique.

While many methods addressing debiasing language models focus on the bias present in the model itself, our empirically-based method offers stronger debiasing of the data at hand. Modern language models are capable of detecting gender signals in a wide variety of datasets ranging from conversational to highly technical language. Many methods for removing bias from the pre-trained language model still leave the potential of meaningful proxies in the target dataset, while also raising questions on degradation in performance. We believe that balancing debiasing with model performance is benefited by interpretable techniques, such as those we present here. For instance, our bias ranking and iterative application of divergence measures allow users to get a sense of disagreement in bias ranks for language models and empirical data.

Our study is limited to looking at a) intrinsic bias found in our dataset and pre-trained word embeddings, and b) the extrinsic potential bias identified in our classification pipeline. We recognize concerns raised by Blodgett *et al.* [173] and others relating to the imprecise definitions of bias in the field of algorithmic fairness.

Indeed, it is often important to motivate a given case of bias by establishing its potential harms. In this piece, we address precursors to bias, and thus do not claim to being making robust assessments of real-world bias and subsequent impact. The potential bias metric is instead meant to be a task agnostic indicator of the capacity for a complete pipeline to discriminate between protected classes.

Due to the available data we were not able to develop methods that address non-binary cases of gender bias. There are other methodological considerations for expanding past the binary cases [181], although this is an important topic for a variety of bias types [182].

There are further complications when moving away from tasks where associated language is not as neatly segmented. For instance, we show above that when evaluating language models such as BERT much of the gendered language largely appears in a readily identifiable region of the semantic space. As a rough heuristic: terms appearing in a medical dictionary tended to be less similar to gendered terms than terms that might appear in casual conversation. For doctors notes, the bulk of the bias stems from words that are largely distinct from those that we expect to be most informative for medically relevant tasks. Further research is required to determine the efficacy of our techniques in domains where language is not as neatly semantically segmented.

Using clinical notes from an ICU context could bias our results due to the types of patients, conditions, and interactions that are common in this setting. For instance, there may be fewer verbal patient-provider interactions reflected in the data and social histories may not be as in-depth (compared with a primary-care setting). Further, the ways in which clinicians code health conditions may vary across contexts, institutions,

and providers. In our study we aim to reduce the impact of how conditions are coded by selecting common conditions that have large sample sizes in our data set—but this is still a factor that should be considered when working with such data.

Future research that applies these interpretable methods to clinical text have the opportunity to examine possible confounding factors such as patient-provider gender concordance. Further, it would be worthwhile to separately address the impact of author gender on the content of clinical texts through using analytical framework. Other confounding factors relating to the patient populations and broader socio-demographic factors could be addressed by replicating these trials on new data sets. There is also the potential to research how presenting the results of our empirical bias analysis to clinicians may affect note writing practices—perhaps adding empirical examples to the growing medical school curriculum that addresses unconscious bias [183].

Our methods make no formal privacy guarantees nor do we claim complete removal of bias. There is always a trade-off when seeking to balance bias reduction with overall performance, and we feel our methods will help all stakeholders make more informed decisions. Our methodology allows stakeholders to specify the trade-off between bias reduction and performance that is best for their particular use case by selecting different trim levels and reviewing the n-grams removed. Using a debiasing method that is readily interpreted by doctors, patients, and machine learning practitioners is a benefit for all involved, especially as public interest in data privacy grows.

Moving towards replacing strings rather than trimming or dropping them completely should be investigated in the future. More advanced data augmentation methods may be needed if we were to explore the impact of debiasing on highly tuned

classification pipelines. Holistic comparisons of string replacement techniques and other text data augmentation approaches would be worthwhile next steps. Further research on varying and more difficult extrinsic evaluation tasks would be helpful in evaluating how our technique generalizes. Future work could also investigate coupling our data-driven method with methods focused on debiasing language models.

3.6 Acknowledgements

The authors are grateful for the computing resources provided by the Vermont Advanced Computing Core and financial support from the Massachusetts Mutual Life Insurance Company and Google. The views expressed are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

3.7 Supplementary Information (SI)

3.7.1 Note selection

After reviewing the note types available in the MIMIC-III dataset, we determined that many types were not suitable for our task. This is due to a combination of factors including information content and note length (Fig. 3.22. Note types such as `radiology` often include very specific information (not indicative of broader patient health status), are shorter, and may be written in a jargonistic fashion. For the work outlined here we only include notes that are of the types `nursing`, `discharge`

`summary`, and `physician`. In order to be included in our training and test datasets, documents must come from patients with at least three recorded documents.

3.7.2 Document lengths after trimming

3.7.3 Variable length note embedding

When tokenized, many of notes available in the MIMIC-III dataset are longer than the 512-token maximum supported by BERT. To address this issue with experiment with truncating the note at the first 512 tokens. We also explore embedding at the sentence level (embedding with a maximum of 128 tokens) and simply dividing the note in 512-token subsequences. In the latter two cases, we use the function outlined by Huang *et al.* [28],

$$P(Y = 1) = \frac{P^n_{max} + P^n_{mean} n/c}{1 + n/c} \tag{3.3}$$

where P^n_{max} and P^n_{mean} are the maximum and mean probabilities for the n subsquences associated with a given note. Here, c is a tunable parameter that is adjusted for each task.

For our purposes, the improvement in classification performance returned by employing this technique did not merit use in our final results. If overall performance of our classification system were our primary objective, this may be worth further investigation.

3.7.4 ICD Co-occurrence

3.7.5 Hardware

BERT and Clinical BERT models were fine-tuned on both an NVIDIA RTX 2070 (8GB VRAM) and NVIDIA Tesla V100s (32GB VRAM).

3.7.6 Gendered 1-grams

Female 1**-grams**	**Male** 1**-grams**
her	his
she	he
woman	man
female	male
Ms	Mr
Mrs	him
herself	himself
girl	boy
lady	gentleman

Table 3.4: ***Manually selected gendered terms.***

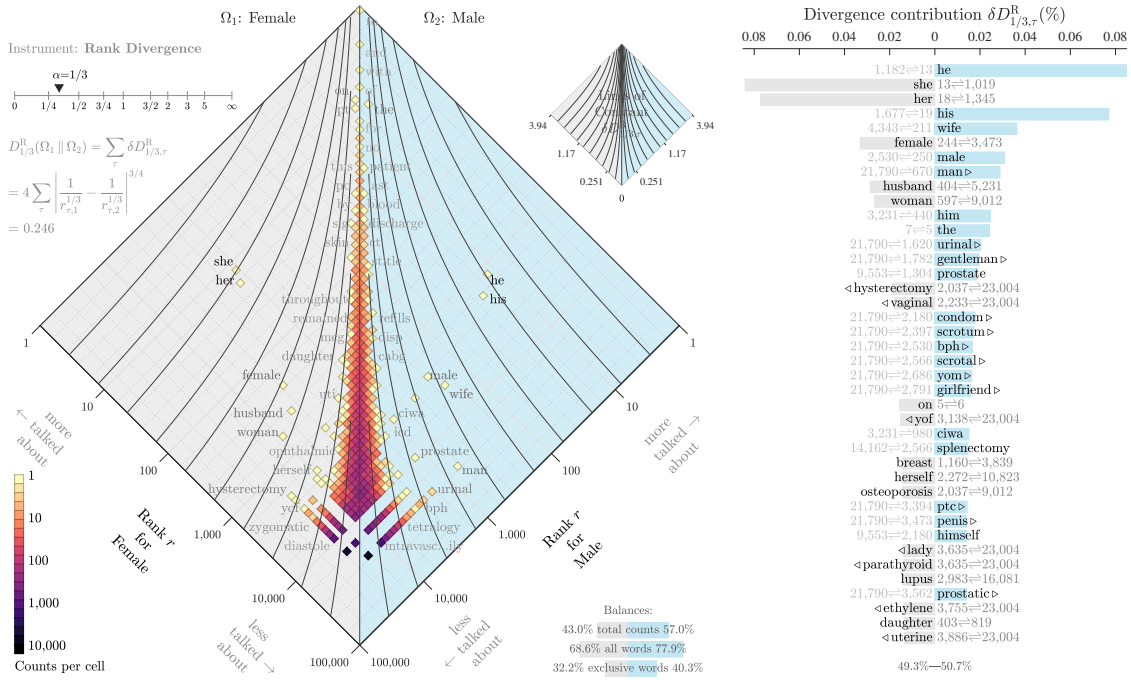


Figure 3.1: ***Rank-turbulence divergence allotaxonograph*** *[3]* ***for male and female documents in the MIMIC-III dataset.*** *For this figure, we generated 1-gram frequency and rank distributions from documents corresponding to male and female patients. Pronouns such as "she" and "he" are immediately apparent as drivers of divergence between the two corpora. From there, the histogram on the right highlights gendered language that is both common and medical in nature. Familial relations (e.g., "husband" and "daughter") often present as highly gendered according to our measure. Further, medical terms like "hysterectomy" and "scrotum" are also highly ranked in terms of their divergence. Higher divergence contribution values, $\delta D^{\mathrm{R}}_{\alpha,\tau}$, are often driven by either relatively common words fluctuating between distributions (e.g., "daughter"), or the presence of disjoint terms that appear in only one distribution (e.g., "hysterectomy"). The impact of higher rank values can be tuned by adjusting the α parameter. In the main horizontal bar chart, the bars indicate the divergence contribution value and the numbers next to the terms represent their rank in each corpus. The terms that appear in only one corpus are indicated with a rotated triangle. The smaller three vertical bars describe balances between the male and female corpora: 43% of total 1gram counts appear in the female corpus; we observed 68.6% of all 1grams in the female corpus; and 32.2% of the 1grams in the female corpus are unique to that corpus.*

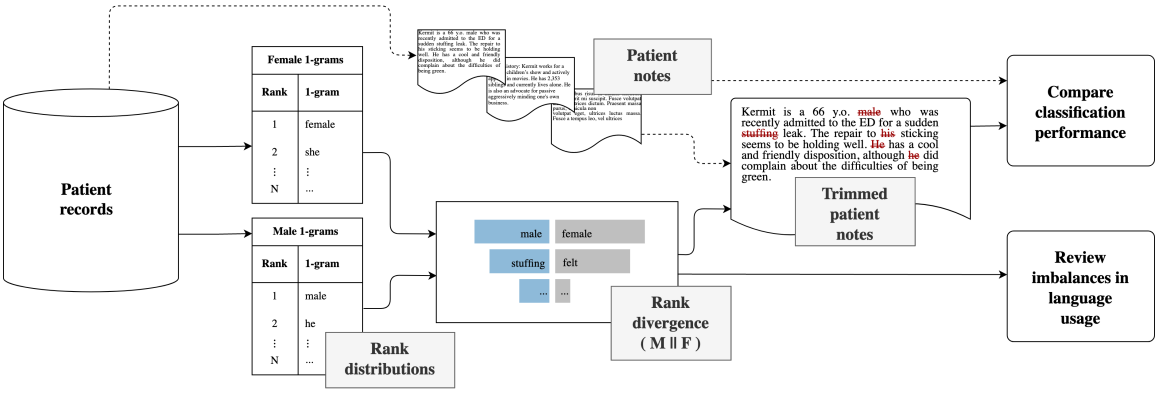


Figure 3.2: ***Overview of the rank-turbulence divergence trimming procedure.*** *Solid lines indicate steps that are specific to our trimming procedure and evaluation process. The pipeline starts with a repository of patient records that include clinical notes and class labels (in our case gender and ICD9 codes). From these notes we generate n-gram rank distributions for the female and male patient populations, which are then used to calculate the rank-turbulence divergence (RTD) for individual n-grams. Sorting the n-grams based on RTD contribution, we then trim the clinical notes. Finally, we view the results directly from the RTD calculation to review imbalance in language use. With the trimmed documents we compare the performance of classifiers on both the un-trimmed notes and notes with varying levels of trimming applied.*

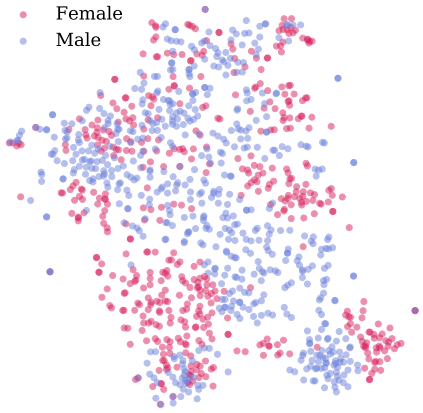


Figure 3.3: ***A tSNE embedding of n2c2 document vectors generated using a pre-trained version of BERT with off-the-shelf weights.*** *We observe the appearance of gendered clusters even before training for a gender classification task. See Fig. 3.17 for the same visualization but with Clinical BERT embeddings.*

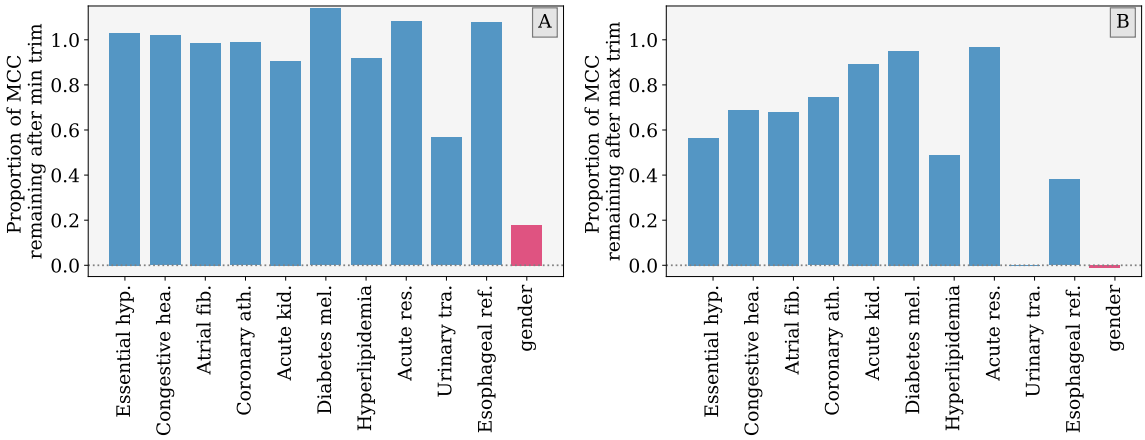


Figure 3.4: ***Patient condition and gender classification performance.*** *Using the fine-tuned Clinical BERT based model on the MIMIC dataset. (**A**) Proportion of baseline classification performance removed after minimum-trim level (1% of total RTD) is applied to the documents. (**B**) Same as (**A**) but with maximum trimming applied (70% of total RTD). Of all the classification tasks, 'gender' and 'Urinary tra.' experience the greatest relative decrease in classification performance. However, due to the low baseline performance of Urinary (≈ 0.2), the gender classification task has a notably higher absolute reduction in MCC than Urinary tra. (or any other task). It is worth noting under low levels of trimming MCC values slightly improved in individual trials. Further, under maximum trim levels the gender classification MCC was slightly negative. See Fig. 3.5 for full information on MCC scores for each of the health conditions.*

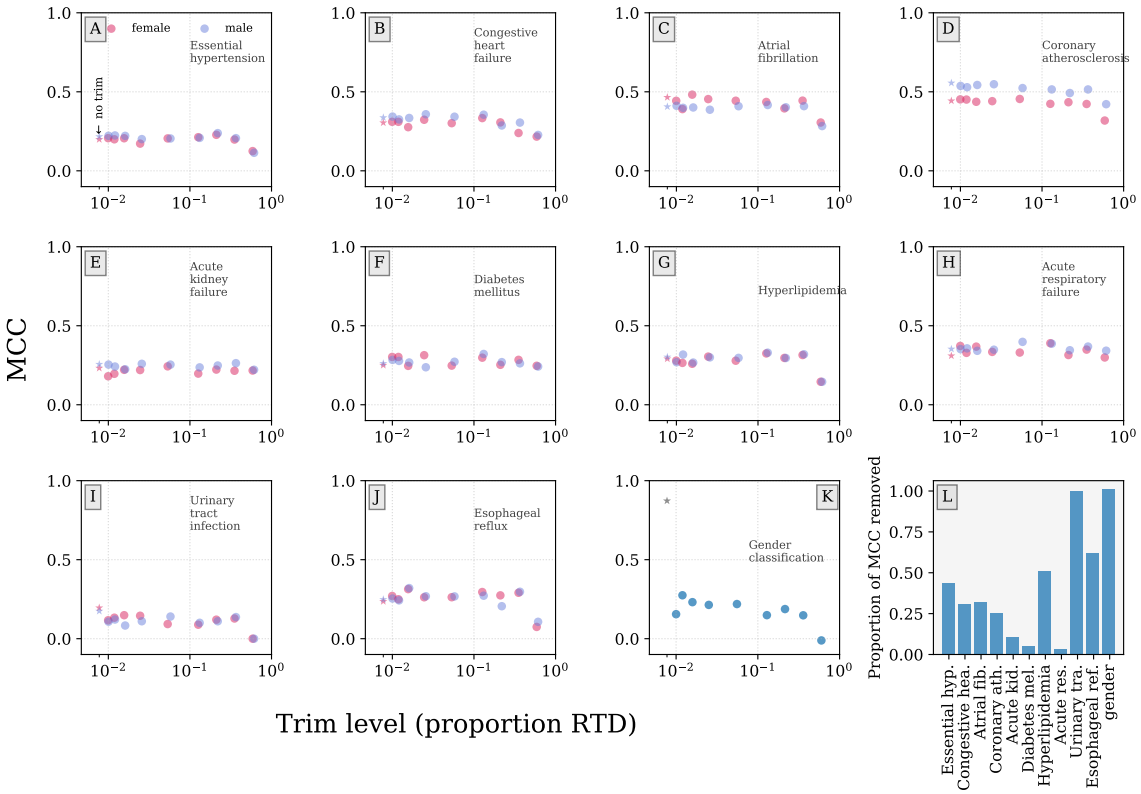


Figure 3.5: ***Matthews correlation coefficient (MCC) for classification results of health conditions and patient gender with varying trim levels****. Results were produced with clinicalBERT embeddings and no-token n-gram trimming. (**A**) – (**J**) show MCC for the top 10 ICD9 codes present in the MIMIC data set. (**K**) shows MCC for gender classification on the same population. (**L**) presents a comparison of MCC results for data with no trimming and the maximum trimming level applied. Values are the relative MCC, or the proportion of the best classifiers performance we lose when applying the maximum rank-turbulence divergence trimming to the data. Here we see the relatively small effect of gender-based rank divergence trimming on the condition classification tasks for most conditions. The performance on the gender classification task is significantly degraded, even at modest trim levels, and is effectively no better than random guessing at our maximum trim level. It is worth noting that many conditions are stable for most of the trimming thresholds, although we do start to see more consistent degradation of performance at the maximum trim level for a few conditions.*

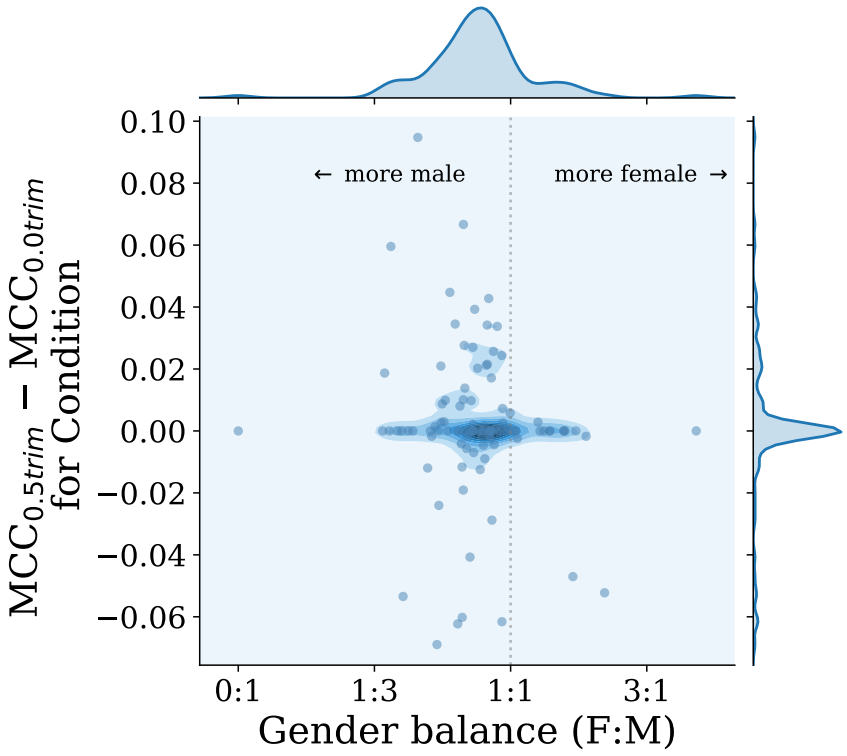


Figure 3.6: ***Degradation in performance for Matthews correlation coefficient for condition classification of ICD9 codes with at least 1000 patients.*** *The performance degradation is presented relative to the proportion of the patients with that code who are female. We find little correlation between the efficacy of the condition classifier on highly augmented (trimmed) datasets and the gender balance for patients with that condition (coefficient of determination $R^2 = -2.48$). Values are calculated for TF-IDF based classifier and include the top 10 health conditions we evaluate elsewhere.*

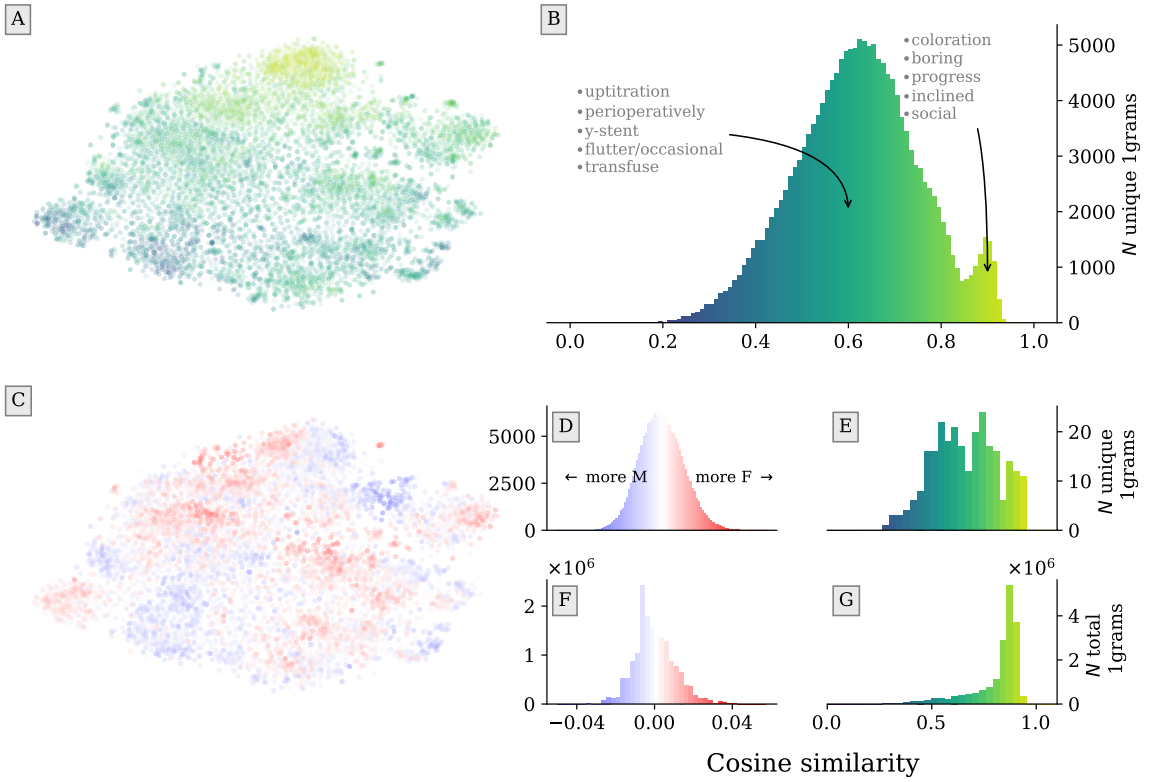


Figure 3.7: ***Measures of gender bias in BERT word-embeddings****. (**A**) tSNE visualization of the BERT embedding space, colored by the maximum cosine similarity of MIMIC-III 1grams to either male or female gendered clusters. (**B**) Distribution of the maximum cosine similarity between male or female gender clusters for 163,539 1grams appearing the MIMIC-III corpus. Through manual inspection we find that the two clusters of cosine similarity values loosely represent more conversational English (around 0.87) and more technical language (around 0.6). The words shown here were manually selected from 20 random draws for each respective region. (**C**) tSNE visualization of BERT embeddings space, colored by the difference in the values of cosine similarity for each word and the male and female clusters. (**D**) Distribution of the differences in cosine similarity values for 1-grams and male and female clusters. (**E**) Distribution maximum gendered-cluster cosine similarity scores for the 1-grams selected for removal when using the rank-turbulence divergence trim technique and targeting the top 1% of words that contribute to overall divergence. The trimming procedure targets both common words that are considered relatively gendered by the cosine similarity measure, and less common words that are more specific to the MIMIC-III dataset and relatively less gendered according to the cosine similarity measure. (**F**) Weighted distribution of differences in cosine similarity between 1-grams and male and female clusters (same measure as (**D**), but weighted by the total number of occurrences of the 1-gram in the MIMIC-III data). (**G**) Weighted distribution of maximum cosine similarity scores between 1-grams and male or female clusters (same measure as (**B**), but weighted by the total number of occurrences of the 1-gram in the MIMIC-III data).*

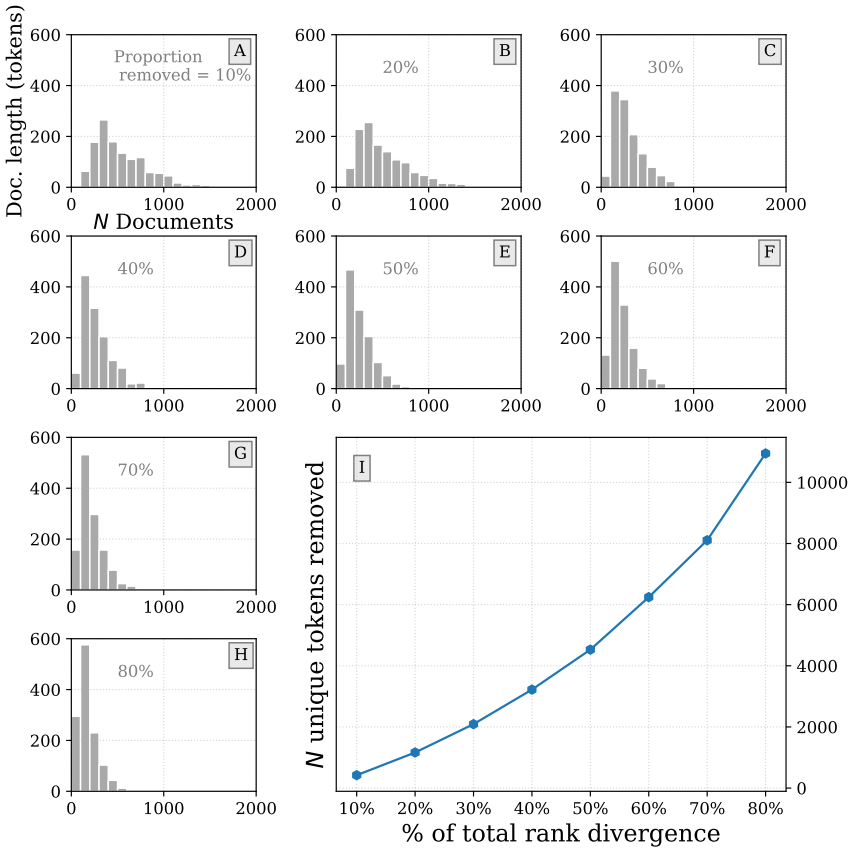


Figure 3.8: ***Document length after applying a linearly-spaced rank-turbulence divergence based trimming procedure.*** *Percentage values represent the percentage of total rank-turbulence divergence removed. Trimming is conducted by sorting words highest-to-lowest based on their individual contribution to the rank-turbulence divergence between male and female corpora (i.e., the first 10% trim will include words that, for most distributions, contribute far more to rank-turbulence divergence than the last 10%).*

Figure 3.9: ***Normalized rates of health-condition co-occurrence for the top 10 ICD-9 codes.***

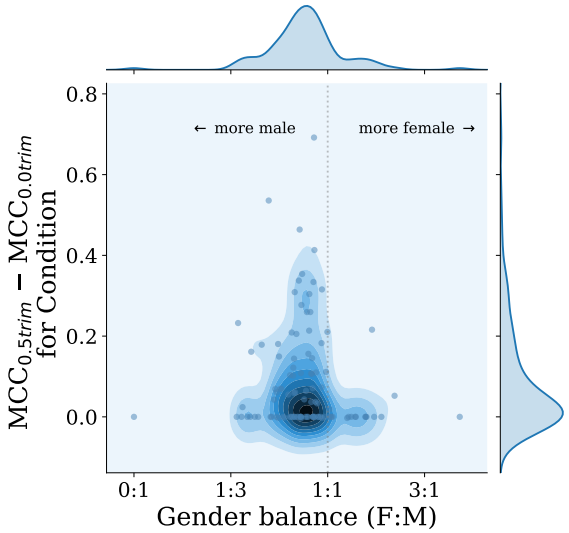


Figure 3.10: ***Classification performance for the next 123 most frequently occurring conditions****. Matthews correlation coefficient for condition classification of ICD9 codes with at least 1000 patients compared to the proportion of the patients with that code who are female. While the most accurate classifiers tend to be for conditions with a male bias, we observed that this is in-part due to the underlying bias in patient gender.*

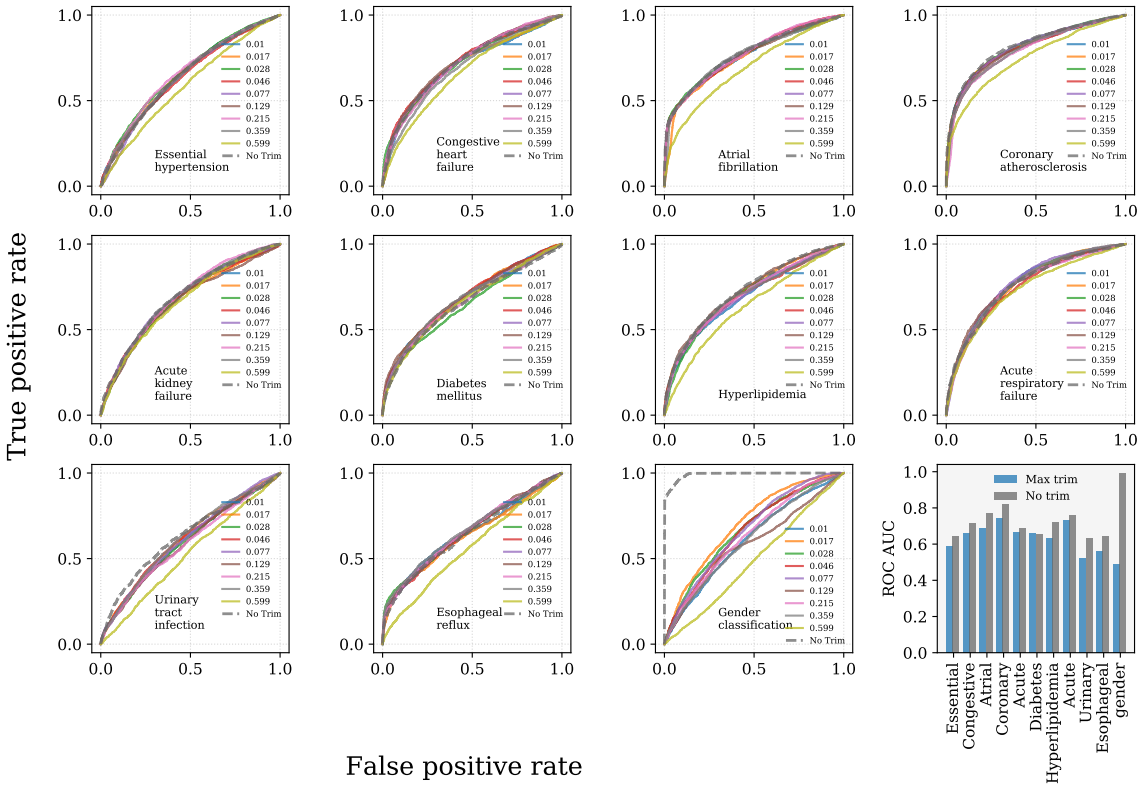


Figure 3.11: ***ROC curves for classification task on top 10 health conditions with varying proportions of rank-turbulence divergence removed.*** *Echoing the results in Fig. 3.5, the gender classifier has the best performance on the 'no-trim' data and experiences the greatest drop in performance when trimming in applied. Under the highest trim level reported here, the gender classifier is effectively random, while few condition classifiers retain prediction capability (albeit modest). The bar chart show the area under the ROC curve for classifiers, by task, trained and tested with no-trimming and maximum-trimming applied.*

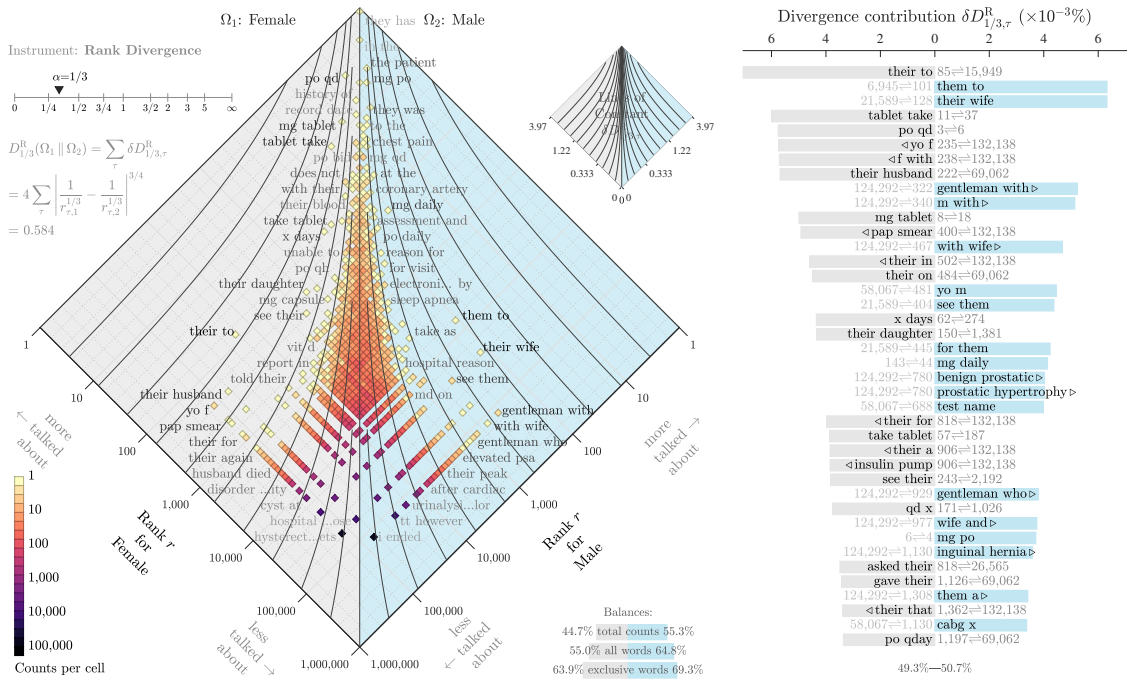


Figure 3.12: ***Rank-turbulence divergence for 2014 n2c2 challenge.*** *For this figure, 2-grams have been split between genders and common gendered terms (pronouns, etc.) have been removed before calculating rank divergence.*

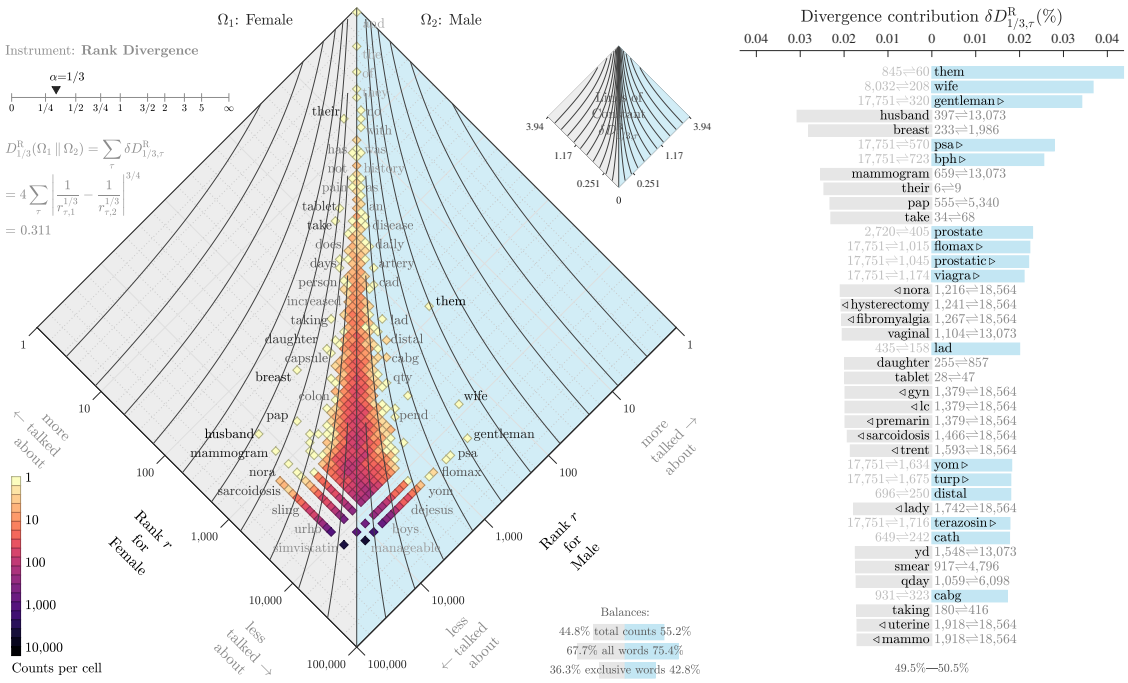


Figure 3.13: ***Rank-turbulence divergence for 2014 n2c2 challenge.*** *For this figure, 1-grams have been split between genders and common gendered terms (pronouns, etc. see Table 3.4) have been removed before calculating rank divergence.*

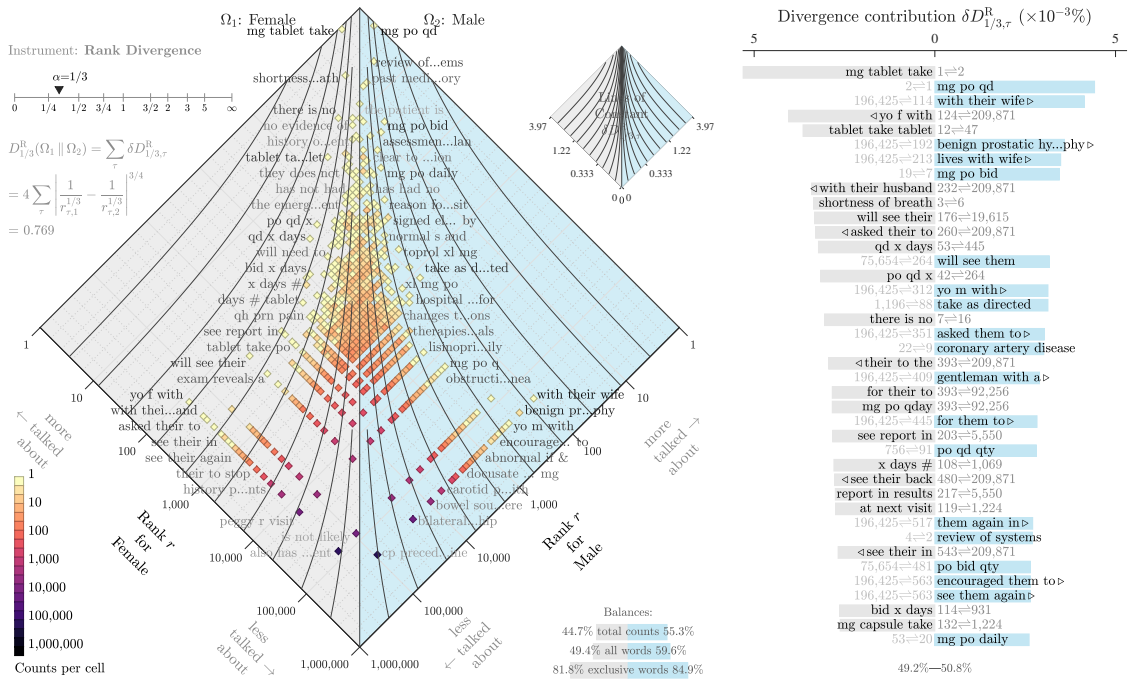


Figure 3.14: ***Rank-turbulence divergence for 2014 n2c2 challenge.*** *For this figure, 3-grams have been split between genders and common gendered terms (pronouns, etc.) have been removed before calculating rank divergence.*

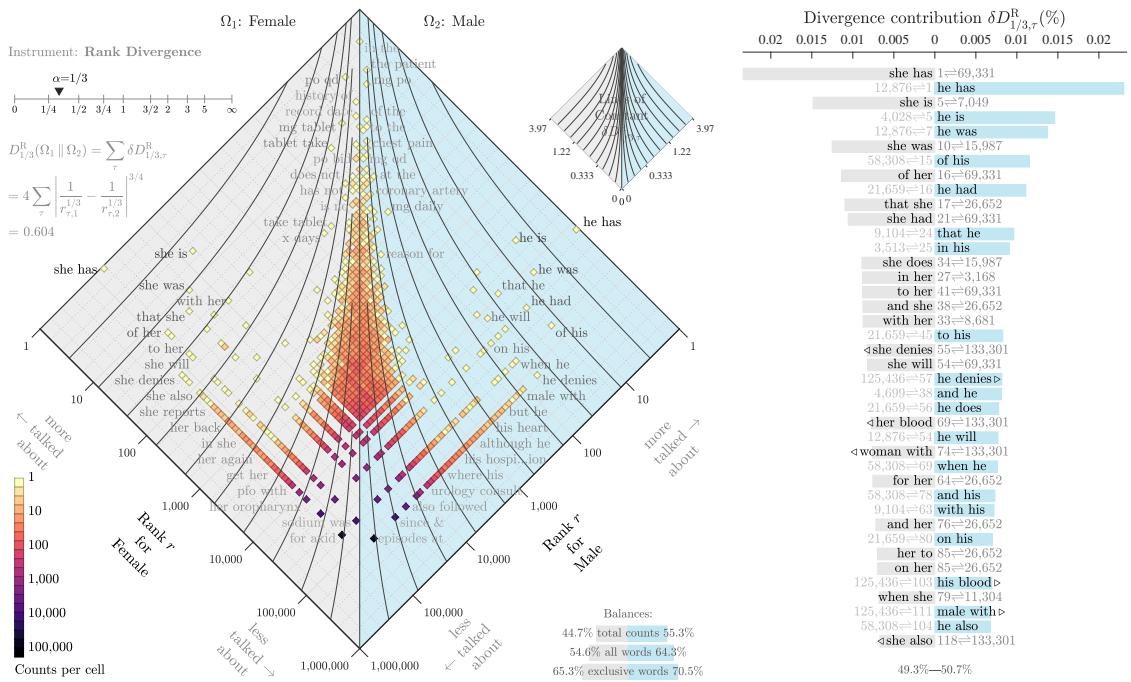


Figure 3.15: ***Rank-turbulence divergence for 2014 n2c2 challenge.*** *For this figure, 2-grams have been split between genders.*

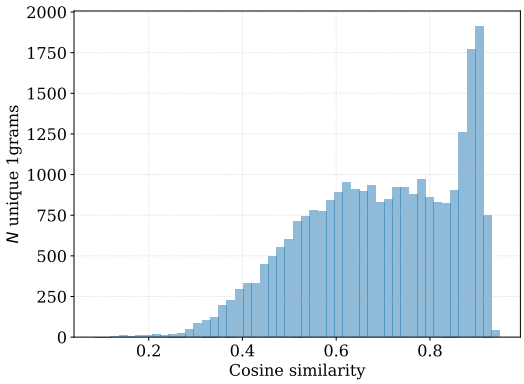


Figure 3.16: Maximum cosine similarity scores of BERT-base embeddings for 26,883 1grams appearing in n2c2 2014 challenge data relative to gendered clusters.

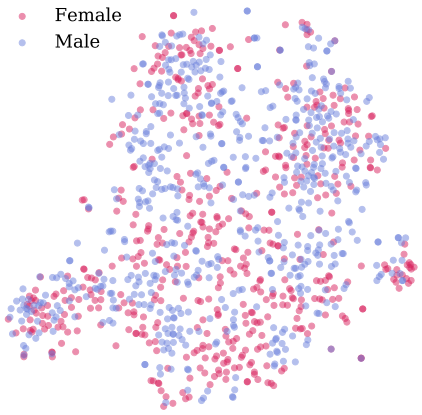


Figure 3.17: A tSNE embedding of n2c2 document vectors generated using a pre-trained version of Clinical BERT.

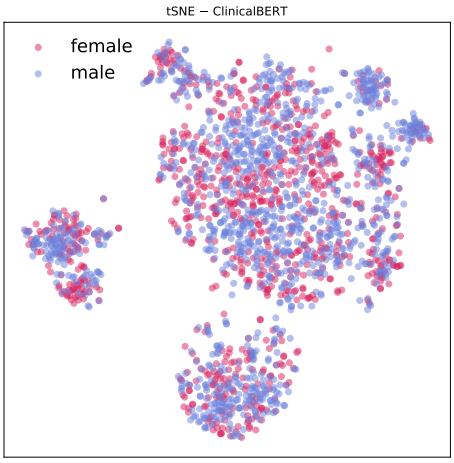


Figure 3.18: ***A tSNE embedding of MIMIC document vectors generated using a pre-trained version of Clinical BERT.***

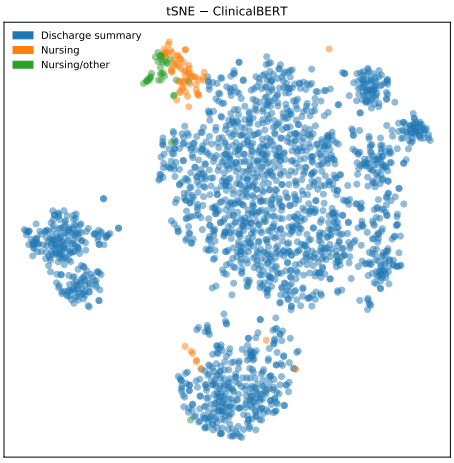


Figure 3.19: ***A tSNE embedding of MIMIC document vectors generated using a pre-trained version of Clinical BERT.***

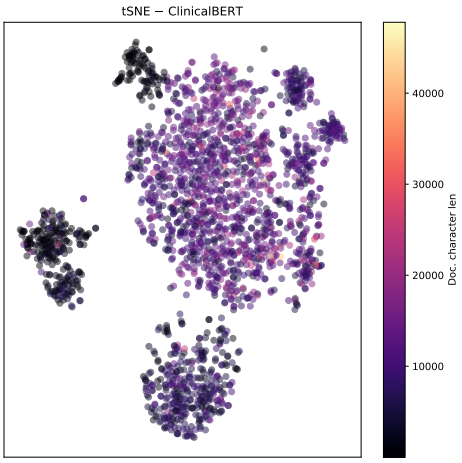


Figure 3.20: ***A tSNE embedding of MIMIC document vectors generated using a pre-trained version of Clinical BERT.***

	1gram	BERT RTD rank	n2c2 RTD rank	BERT-n2c2 RTD rank
1	mrs	1.0	1172.5	1.0
2	ms	4.0	6560.5	2.0
3	her	279.0	3.0	3.0
4	mr.	15307.0	6.0	4.0
5	male	21798.0	8.0	5.0
6	mr	5.0	437.0	6.0
7	female	5150.0	9.0	7.0
8	linda	10.0	3681.5	8.0
9	ms.	2208.0	11.0	9.0
10	gentleman	3054.0	13.0	10.0
11	pap	25105.0	17.0	11.0
12	breast	4314.0	14.0	12.0
13	cervical	14.0	2483.0	13.0
14	biggest	16.0	5301.5	14.0
15	mammogram	25054.0	21.0	15.0
16	mrs.	2860.0	16.0	16.0
17	f	10458.0	20.0	17.0
18	woman	120.0	7.0	18.0
19	psa	6082.0	19.0	19.0
20	kathy	19.0	5301.5	20.0
21	he	3.0	1.0	21.0
22	them	20.0	4146.0	22.0
23	prostate	2223.0	18.0	23.0
24	bph	8601.0	23.0	24.0
25	husband	920.0	15.0	25.0
26	guy	22.0	5301.5	26.0
27	take	4278.0	22.0	27.0
28	infected	18.0	1455.0	28.0
29	patricia	21.0	2701.5	29.0
30	smear	21455.0	29.0	30.0
31	ellen	24.0	3681.5	31.0
32	cabg	19064.0	33.0	32.0
33	distal	7754.0	30.0	33.0
34	pend	22515.0	36.0	34.0
35	tablet	2322.0	25.0	35.0
36	cath	7111.0	31.0	36.0
37	qday	12829.0	34.0	37.0
38	peggy	17.0	485.0	38.0
39	flomax	17413.0	37.5	39.0
40	lad	2383.0	27.0	40.0
41	prostatic	23978.0	43.0	41.0
42	gout	11948.0	40.0	42.0
43	taking	9534.0	39.0	43.0
44	trouble	34.0	4183.5	44.0
45	harry	33.0	3372.0	45.0

1gram	BERT RTD rank	MIMIC RTD rank	BERT-MIMC RTD rank	MIMIC F rank	MIMIC M rank
sexually	2.0	16755.5	1.0	10607.0	10373.5
biggest	3.0	7594.5	2.0	17520.5	21172.5
f	6288.0	3.0	3.0	251.0	1719.0
infected	5.0	16076.0	4.0	3475.0	3402.5
grandmother	4.0	3517.0	5.0	6090.0	7643.0
cervical	7.0	18374.0	6.0	1554.0	1551.5
m	244.0	2.0	7.0	2103.0	249.0
sister	6.0	4153.0	8.0	1213.5	1365.0
husband	860.0	4.0	9.0	495.0	5114.0
teenage	9.0	5513.0	10.0	15449.0	19594.0
trouble	12.0	10119.0	11.0	3778.5	4095.5
brother	10.0	1921.0	12.0	1925.0	1598.0
teenager	8.0	936.5	13.0	12928.5	21172.5
connected	16.0	16682.0	14.0	5184.5	5089.5
shaky	11.0	2341.0	15.0	10607.0	7872.0
my	15.0	8198.0	16.0	2395.0	2624.5
breast	3397.0	6.0	17.0	1075.0	4673.5
expelled	19.0	14652.5	18.0	20814.0	19594.0
them	18.0	11337.0	19.0	1738.5	1832.0
prostate	2196.0	5.0	20.0	9436.0	1576.5
immune	20.0	15043.5	21.0	9119.0	8753.0
daughter	1.0	16.0	22.0	463.0	801.5
initial	23.0	11433.0	23.0	632.5	610.0
ovarian	16374.0	8.0	24.0	3137.0	14082.0
recovering	24.0	11867.0	25.0	5351.5	5749.5
abnormal	25.0	9010.0	26.0	1849.5	1994.0
alcoholic	17.0	1136.0	27.0	3885.5	2952.0
obvious	26.0	10154.0	28.0	2725.5	2540.5
huge	28.0	13683.5	29.0	8069.0	7643.0
dirty	29.0	13264.5	30.0	8613.5	9179.5
suv	27.0	5911.0	31.0	14704.0	18377.5
container	31.0	17776.0	32.0	12090.5	12214.5
flomax	10891.0	11.0	33.0	17520.5	4095.5
sisters	32.0	17680.5	34.0	4727.0	4687.5
uterine	7260.0	10.0	35.0	4263.0	19594.0
hypothyroidism	17831.0	12.0	36.0	1239.5	2920.0
dried	34.0	15661.0	37.0	5124.5	4982.0
osteoporosis	18010.0	13.0	38.0	2354.0	7003.0
breasts	4727.0	9.0	39.0	4394.5	21172.5
certain	37.0	18020.5	40.0	7973.0	8023.0
i	30.0	4601.0	41.0	403.0	435.5
restless	21.0	1144.0	42.0	1366.0	1123.0
wife	13.0	1.0	43.0	5545.0	245.5
sle	8083.0	14.0	44.0	3511.0	12504.5
granddaughter	14.0	210.0	45.0	4656.5	7872.0
localized	47.0	14357.5	46.0	6136.0	6405.0
ciwa	7921.0	15.0	47.0	3106.5	1341.0
honey	44.0	11076.0	48.0	10868.5	9861.0
coronary	11570.0	18.0	49.0	560.0	349.0
systemic	41.0	6361.0	50.0	3696.0	4214.0

Table 3.6: ***Comparison of rank-turbulence divergences for gendered clusters in BERT embeddings and the MIMIC patient health records text.*** *BERT RTD ranks are calculated based on cosine similarity scores for word embedding and gendered clusters (i.e., the RTD of cosine similarity score ranks relative to male and female clusters). MIMIC RTD ranks are for* 1*-grams from male and female clinical notes. "BERT-MIMIC RTD rank" is the rankings for* 1*-grams based on RTD between the first two columns—we also refer to this as RTD*^{2} *(ranking divergence-of-divergence).*

1gram	BERT RTD rank	MIMIC RTD rank	BERT-MIMC RTD rank	MIMIC F rank	MIMIC M rank
is	1.0	18545.5	1.0	24.0	24.0
wife	3588.0	1.0	2.0	245.5	5545.0
yells	2.0	11292.0	3.0	10563.0	9615.0
looking	3.0	16700.0	4.0	3238.5	3289.5
m	7181.0	2.0	5.0	249.0	2103.0
kids	4.0	13675.0	6.0	8753.0	9266.5
essentially	5.0	17864.0	7.0	2269.0	2279.5
alter	6.0	15840.0	8.0	15764.0	16387.5
f	2166.0	3.0	9.0	1719.0	251.0
bumps	7.0	6936.0	10.0	13612.0	11417.0
husband	2958.0	4.0	11.0	5114.0	495.0
historian	8.0	7645.0	12.0	6895.0	6045.5
insult	9.0	7241.0	13.0	8854.5	10361.0
moments	10.0	16287.0	14.0	10563.0	10868.5
our	11.0	16938.0	15.0	2803.0	2838.0
asks	13.0	16147.0	16.0	7257.0	7449.5
goes	15.0	17662.0	17.0	3653.0	3624.5
someone	16.0	14030.0	18.0	6197.0	5920.5
ever	17.0	10698.0	19.0	5114.0	5545.0
prostate	6070.0	5.0	20.0	1576.5	9436.0
breast	6010.0	6.0	21.0	4673.5	1075.0
experiences	20.0	17062.0	22.0	9452.5	9615.0
suffer	12.0	1463.5	23.0	10968.5	16387.5
recordings	21.0	13249.5	24.0	12504.5	13440.5
wore	23.0	14774.5	25.0	8854.5	9266.5
largely	26.0	16941.0	26.0	4581.5	4515.5
hi	22.0	6459.0	27.0	3992.5	4558.0
et	27.0	17615.0	28.0	2079.0	2093.5
staying	25.0	10254.5	29.0	4366.5	4746.5
pursuing	18.0	1561.5	30.0	13612.0	20814.0
pet	24.0	4873.0	31.0	5879.0	7076.5
town	28.0	10038.0	32.0	10754.0	9615.0
tire	30.0	12053.5	33.0	12834.5	11736.5
ovarian	13390.0	8.0	34.0	14082.0	3137.0
beef	19.0	1355.5	35.0	23307.0	14704.0
dipping	34.0	17764.0	36.0	6371.0	6318.5
dip	35.0	18328.0	37.0	5585.0	5600.0
hat	33.0	10106.5	38.0	14082.0	12478.5
flomax	14782.0	11.0	39.0	4095.5	17520.5
punch	31.0	5064.5	40.0	11203.5	14033.0
ease	38.0	17265.5	41.0	6471.5	6556.0
hasn	36.0	11324.0	42.0	10373.5	11417.0
lasts	39.0	14072.5	43.0	11203.5	10607.0
grabbing	32.0	4294.0	44.0	10968.5	14033.0
hypothyroidism	17820.0	12.0	45.0	2920.0	1239.5
whatever	37.0	8524.5	46.0	13200.5	15449.0
osteoporosis	18276.0	13.0	47.0	7003.0	2354.0
uterine	6519.0	10.0	48.0	19594.0	4263.0
sle	16374.0	14.0	49.0	12504.5	3511.0
dump	47.0	15439.0	50.0	10968.5	11417.0

Table 3.7: ***Comparison of rank-turbulence divergences for gendered clusters in Clinical BERT embeddings and the MIMIC patient health records text.*** *Clinical BERT RTD ranks are calculated based on cosine similarity scores for word embedding and gendered clusters (i.e., the RTD of cosine similarity score ranks relative to male and female clusters). MIMIC RTD ranks are for* 1*-grams from male and female clinical notes. "BERT-MIMIC RTD rank" is the rankings for* 1*-grams based on RTD between the first two columns—we also refer to this as* RTD^2 *(ranking divergence-of-divergence). The presence of largely conversational terms rather than more technical, medical language owes to our defining of gender clusters through manually selected terms.*

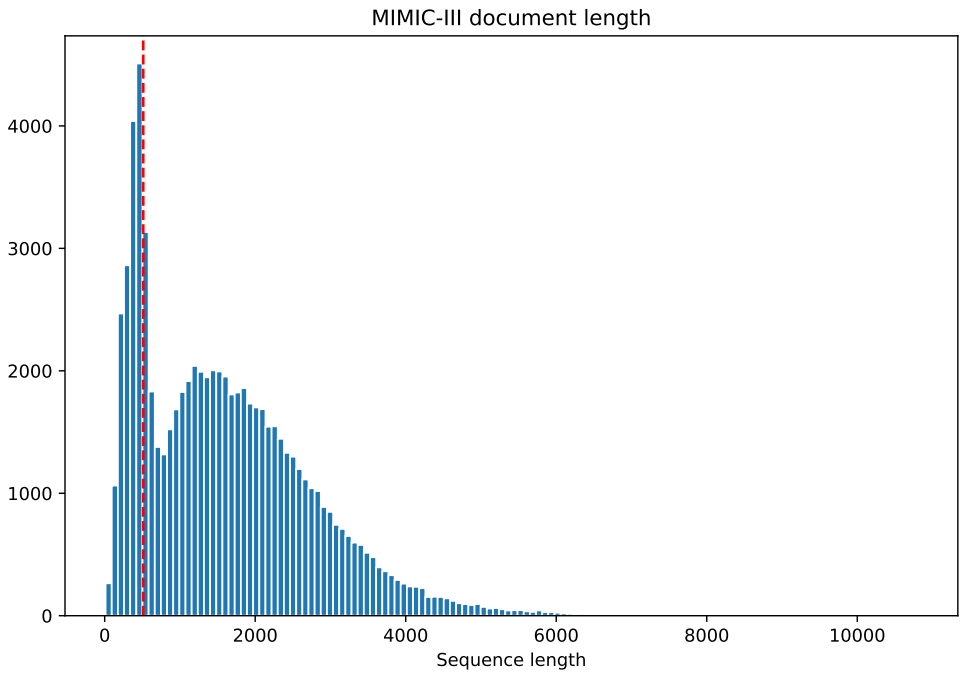


Figure 3.21: ***Document length for MIMIC-III text notes.***

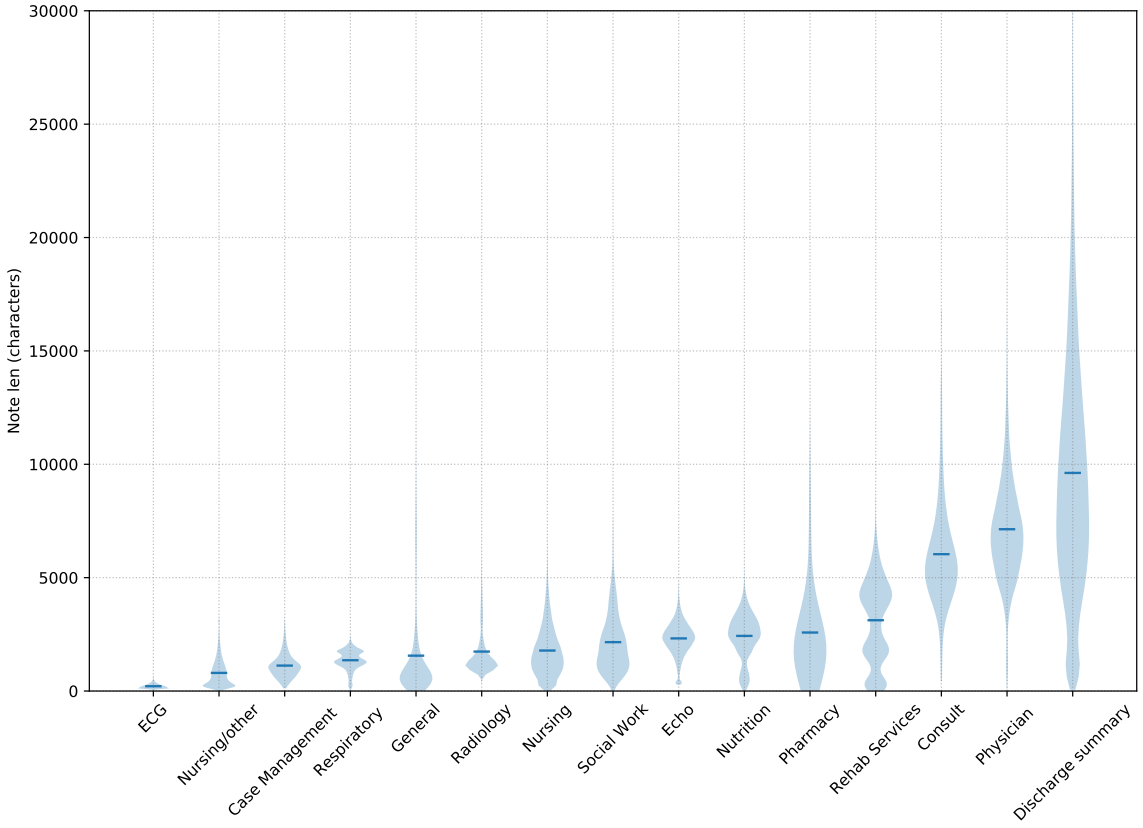


Figure 3.22: ***Document length for MIMIC-III by note type.*** *For our study we include discharge summary, physician, and nursing notes. Consult notes were initially considered but were ultimately found to be highly varied in terms of notation and nomenclature. This had the effect of making results more difficult to interpret and would have required additional data cleaning. We believe our methods could be applied to patient records that include consult notes, just at the cost of additional pre-processing and more nuanced interpretation.*

	Sex Count		Sex Prop.	
ICD Description.	**F**	**M**	**F**	**M**
Personal history of malignant neoplasm of prostate	0	1207	0.00	1.00
Hypertrophy (benign) of prostate without urinar...	0	1490	0.00	1.00
Routine or ritual circumcision	0	2016	0.00	1.00
Gout, unspecified	552	1530	0.27	0.73
Alcoholic cirrhosis of liver	323	879	0.27	0.73
Retention of urine, unspecified	283	737	0.28	0.72
Intermediate coronary syndrome	466	1197	0.28	0.72
Chronic systolic heart failure	321	776	0.29	0.71
Aortocoronary bypass status	896	2160	0.29	0.71
Other and unspecified angina pectoris	330	770	0.30	0.70
Paroxysmal ventricular tachycardia	548	1263	0.30	0.70
Chronic hepatitis C without mention of hepatic ...	380	838	0.31	0.69
Coronary atherosclerosis of unspecified type of...	479	1015	0.32	0.68
Percutaneous transluminal coronary angioplasty ...	889	1836	0.33	0.67
Portal hypertension	332	675	0.33	0.67
Surgical operation with anastomosis, bypass, or...	406	805	0.34	0.66
Coronary atherosclerosis of native coronary artery	4322	8107	0.35	0.65
Old myocardial infarction	1156	2122	0.35	0.65
Acute on chronic systolic heart failure	406	737	0.36	0.64
Cardiac complications, not elsewhere classified	847	1496	0.36	0.64
Atrial flutter	444	773	0.36	0.64
Paralytic ileus	394	678	0.37	0.63

	Sex Count		Sex Prop.	
ICD Description.	**F**	**M**	**F**	**M**
Chronic kidney disease, unspecified	1265	2170	0.37	0.63
Personal history of tobacco use	1042	1769	0.37	0.63
Pneumonitis due to inhalation of food or vomitus	1369	2311	0.37	0.63
Tobacco use disorder	1251	2107	0.37	0.63
Obstructive sleep apnea (adult)(pediatric)	891	1489	0.37	0.63
Cirrhosis of liver without mention of alcohol	486	801	0.38	0.62
Hypertensive chronic kidney disease, unspecifie...	1300	2121	0.38	0.62
Diabetes with neurological manifestations, type...	438	700	0.38	0.62
Other primary cardiomyopathies	664	1045	0.39	0.61
Cardiac arrest	542	819	0.40	0.60
Peripheral vascular disease, unspecified	564	837	0.40	0.60
Hyperpotassemia	874	1295	0.40	0.60
Bacteremia	599	879	0.41	0.59
Other and unspecified hyperlipidemia	3537	5153	0.41	0.59
Thrombocytopenia, unspecified	1255	1810	0.41	0.59
Pure hypercholesterolemia	2436	3494	0.41	0.59
Pressure ulcer, lower back	530	759	0.41	0.59
Subendocardial infarction, initial episode of care	1262	1793	0.41	0.59
Acute kidney failure with lesion of tubular nec...	945	1342	0.41	0.59
Acute and subacute necrosis of liver	441	626	0.41	0.59
Hypertensive chronic kidney disease, unspecifie...	1091	1539	0.41	0.59
Hemorrhage complicating a procedure	637	898	0.41	0.59

	Sex Count		Sex Prop.	
ICD Description.	**F**	**M**	**F**	**M**
Cardiogenic shock	480	674	0.42	0.58
Aortic valve disorders	1069	1481	0.42	0.58
Polyneuropathy in diabetes	667	917	0.42	0.58
Other postoperative infection	503	683	0.42	0.58
Respiratory distress syndrome in newborn	559	755	0.43	0.57
Cardiac pacemaker in situ	592	798	0.43	0.57
Atrial fibrillation	5512	7379	0.43	0.57
Pulmonary collapse	931	1234	0.43	0.57
Delirium due to conditions classified elsewhere	622	823	0.43	0.57
Diabetes mellitus without mention of complicati...	3902	5156	0.43	0.57
Hemorrhage of gastrointestinal tract, unspecified	602	795	0.43	0.57
Other and unspecified coagulation defects	438	578	0.43	0.57
Acute kidney failure, unspecified	3941	5178	0.43	0.57
End stage renal disease	836	1090	0.43	0.57
Accidents occurring in residential institution	456	583	0.44	0.56
Single liveborn, born in hospital, delivered by...	1220	1538	0.44	0.56
Sepsis	563	709	0.44	0.56
Hyperosmolality and/or hypernatremia	1009	1263	0.44	0.56
Other specified surgical operations and procedu...	600	750	0.44	0.56
Severe sepsis	1746	2166	0.45	0.55
Unspecified protein-calorie malnutrition	562	697	0.45	0.55
Long-term (current) use of insulin	1138	1400	0.45	0.55

	Sex Count		Sex Prop.	
ICD Description.	**F**	**M**	**F**	**M**
Long-term (current) use of anticoagulants	1709	2097	0.45	0.55
Other iatrogenic hypotension	953	1168	0.45	0.55
Anemia in chronic kidney disease	623	761	0.45	0.55
Intracerebral hemorrhage	618	749	0.45	0.55
Unspecified essential hypertension	9370	11333	0.45	0.55
Acute posthemorrhagic anemia	2072	2480	0.46	0.54
Unspecified septicemia	1702	2023	0.46	0.54
Chronic airway obstruction, not elsewhere class...	2027	2404	0.46	0.54
Pneumonia, organism unspecified	2223	2616	0.46	0.54
Septic shock	1189	1397	0.46	0.54
Other convulsions	892	1042	0.46	0.54
Other specified procedures as the cause of abno...	693	809	0.46	0.54
Diarrhea	484	565	0.46	0.54
Hematoma complicating a procedure	566	658	0.46	0.54
Acute respiratory failure	3473	4024	0.46	0.54
Other specified cardiac dysrhythmias	1137	1316	0.46	0.54
Need for prophylactic vaccination and inoculati...	2680	3099	0.46	0.54
Personal history of transient ischemic attack (...	498	574	0.46	0.54
Neonatal jaundice associated with preterm delivery	1052	1212	0.46	0.54
Observation for suspected infectious condition	2570	2949	0.47	0.53
Congestive heart failure, unspecified	6106	7005	0.47	0.53
Hypovolemia hyponatremia	641	733	0.47	0.53

	Sex Count		Sex Prop.	
ICD Description.	**F**	**M**	**F**	**M**
Single liveborn, born in hospital, delivered wi...	1668	1898	0.47	0.53
Unspecified pleural effusion	1281	1453	0.47	0.53
Acidosis	2127	2401	0.47	0.53
Esophageal reflux	2990	3336	0.47	0.53
Encounter for palliative care	485	535	0.48	0.52
Hyposmolality and/or hyponatremia	1445	1594	0.48	0.52
Iron deficiency anemia secondary to blood loss ...	482	530	0.48	0.52
Hypoxemia	625	673	0.48	0.52
Mitral valve disorders	1416	1510	0.48	0.52
Primary apnea of newborn	506	537	0.49	0.51
Hypotension, unspecified	996	1055	0.49	0.51
Personal history of venous thrombosis and embolism	786	826	0.49	0.51
Obesity, unspecified	744	767	0.49	0.51
Intestinal infection due to Clostridium difficile	716	728	0.50	0.50
Obstructive chronic bronchitis with (acute) exa...	598	600	0.50	0.50
Anemia of other chronic disease	550	543	0.50	0.50
Anemia, unspecified	2729	2677	0.50	0.50
Dehydration	704	681	0.51	0.49
Other chronic pulmonary heart diseases	1101	1047	0.51	0.49
Do not resuscitate status	694	633	0.52	0.48
Depressive disorder, not elsewhere classified	1888	1543	0.55	0.45
Morbid obesity	648	522	0.55	0.45

	Sex Count		Sex Prop.	
ICD Description.	**F**	**M**	**F**	**M**
Iron deficiency anemia, unspecified	657	514	0.56	0.44
Chronic diastolic heart failure	708	532	0.57	0.43
Hypopotassemia	816	609	0.57	0.43
Anxiety state, unspecified	944	636	0.60	0.40
Dysthymic disorder	663	446	0.60	0.40
Asthma, unspecified type, unspecified	1317	878	0.60	0.40
Urinary tract infection, site not specified	4027	2528	0.61	0.39
Other persistent mental disorders due to condit...	698	428	0.62	0.38
Acute on chronic diastolic heart failure	779	441	0.64	0.36
Unspecified acquired hypothyroidism	3307	1610	0.67	0.33
Osteoporosis, unspecified	1637	310	0.84	0.16
Personal history of malignant neoplasm of breast	1259	18	0.99	0.01

Table 3.8: ***Condition name and gender balance for the ICD9 codes with at least 1000 observations in the MIMIC-III dataset.***

Chapter 4

Gender biases in resume text data

4.1 Abstract

The gender wage gap has remained steady over the past decade with women earning on average 84 cents for every dollar earned by men. The reasons for this are multifaceted and only partially understood. Despite gains in education and workforce representation, there remain challenges related to broad societal factors and workplace gender-discrimination. In the present study we analyze the text from millions of resumes to investigate the extent to which language features are associated with labor market gender disparities such as wages and gender representation. We start by describing observed differences in language distributions along gendered lines—highlighting the gender disparities of sub-specialities within occupations. We go on to describe the resume text space with topic models, and provide a sense for what terms commonly co-occur in individual resumes. We find that differences in job-specific language distributions explain roughly 11% of the variation in gender pay gap on their own. Through the same analysis we find that as job-specific language

differences increase, gender pay gap decreases. We go on to control for gender-biases of words themselves, and find that gender bias poorly describes the variance in wage gap. As part of this study we show that gender-bias of word-embeddings for sets of terms appearing in female and male resumes are strongly associated with the gender balance of an occupation but not the gender wage gap. Taken together these results suggests that textual data is a powerful piece of information for improving worker representations. Additionally, the results highlight the necessity of understanding biases of this data—especially as artificial intelligence increasingly comes into contact with hiring decisions.

4.2 Introduction

Written texts such as resumes, cover letters, and job ads are often a central feature of the hiring process. Free text allows for the representation of skills, experience, and responsibilities with nuance and specificity. Job advertisements often have lengthy position descriptions to attract and select for the right talent, while workers (ideally) take time to carefully craft brief narratives describing the most relevant aspects of their work experience. The presence of text opens the possibility for societal biases to manifest in intuitive and non-intuitive manners. In the modern day this bias may come about both from human and machine reviewers in the hiring pipeline. Bias in hiring is a common case study raised when discussing bias in artificial intelligence (AI) systems [184–187], but relatively little is known about the data set bias present with the primary artifact for job seekers: resumes. With resumes being a central piece of information in the hiring process, any bias present in the data could be incor-

porated into trained models while also having an impact on the individual trajectory of applicants.

The National Institute for Standards and Technology (NIST) has been working towards improved standards for quantifying and understanding bias in AI systems [31]. The main sources of bias in an AI system can be roughly attributed to structural, human, and statistical factors. We can conceptualize mitigation as occurring primarily at the data set level [188, 189], at test and evaluation time, and/or with modification of human factors. In all mitigation approaches it is helpful to understand domain specific bias present in the main data artifacts of the field. Ultimately, bias mitigation has ethical and performance implications. Improving representations of individuals in AI pipelines not only improves outcomes with respect to ethical considerations, but also improves our understanding of systems through a scientific lens. Indeed, we hope the current work can serve to reduce the precursors to bias in AI pipelines while also improving how workers are represented in future-of-word and labor economics.

There are four main stages in the hiring process: sourcing, screening, interviewing, and selection [190, 191]. AI may come into contact with resumes at multiple points in the hiring and employment process. For instance, selecting candidates for recruitment based on public profiles and screening candidates based on their resumes. More generally, AI might be used to understand broader labor market dynamics—including prediction of workers' future jobs [192] or modelling the interaction of skills in labor markets [193]. Other work has sought to improve worker representations beyond simple titles—creating a system that can translate positions across companies [194]. The representativeness of resumes and its interaction with AI thus has impacts for workers in the hiring process and broader efforts to understand labor market dynamics.

Prior work has found gender-biases in language models that mirror broader societal-biases on gender representation in specific jobs [45, 46, 195–197]. These biases have been found in recent state-of-the-art natural language generation models [198]. Some of the same studies use gender proportions of names as a point of reference when evaluating language model biases [44]. In one notable case, names and employment intersected, with hiring software used by a major technology company found to disproportionately select for stereotypically male names [186]. Gendered language in job postings has been found to lead to reduced diversity of applicants and longer time to fill a job, with at least one company selling a product to make the text of job ads gender-neutral [199]. While AI has the potential to perpetuate biases and introduce new forms of bias, there are companies claiming to use screening algorithms where bias is mitigated, or even companies marketing mitigation tools for human resources professionals [200].

Linguistic patterns have been shown to vary with demographic factors such as speaker location [201, 202] and gender [203]. Analysis of Facebook messages found significant variations in vocabulary for female and male users, with females using more emotion words and males swearing more often and writing about objects at a higher rate [204]. Similarly, syntactic features have been found to vary between female and male writers [205]. Another study found the written evaluations of medical students contains language that can vary significantly between male and female students [206].

Artificial intelligence systems operating on text are all but guaranteed to contain some human biases which may be harmful, neutral, or simply reflect on-the-ground realities [44]. Seminal work on gender bias in embedding spaces [46] demonstrated how word embeddings encode societal gender biases that largely match the gender

imbalance of employment in given jobs. Follow-on work for more recent language models shows bias in how gender is encoded, partially stemming from imbalances in training data [207]. Word embeddings have been used as a tool for quantifying language bias over time [45]. The gender bias in word embedding spaces correlates with the gender ratings of professions by human reviewers [208]. Gender bias as it relates to job titles is a common theme in fairness work [44, 46], for instance co-reference resolution of job titles and pro-nouns [195]. Sentiment analysis models have been found to contain similar bias, as outlined in Bhaskaran and Bhallamudi [209]. More generally, gender information can improve performance on NLP tasks utilizing word-embeddings—suggesting gender, textual data, and NLP pipelines interact in notable ways [210]. There are proposed systems to mitigate gender bias in word embeddings [211]. These approaches have been criticized for simply reducing bias in known measurements while leaving other forms of bias unaddressed [212]. However, recent work on gender bias and language classification shows that datasets can be augmented to reduce gender signal while maintaining performance on the classification task with count-based models [187] and with neural network language models [213].

Recent work found that in 2020 the median hourly earnings of women were 84% of men's earnings in the US [214]. Despite progress made on the pay gap in the 1980s and 90s, the rate of convergence has decreased in recent years. This comes at a time when women have greatly increased educational attainment, but gender segregation in specific bachelors and doctoral degrees has remained steady since the 90s [31]. Some of the pay gap is explained by factors such as experience, part-time work, and an overrepresenation of women in lower paying jobs. Between the 1980s and the 2010s the proportion of the wage gap explained by variables such as experience,

industry distribution, education, and unionization has decreased—this has left labor economists looking for variables that address the unexplained portion of the gap [215].

A study of US Census data from 1950 to 2000 found that on average fields that increased the proportion of female workers experienced a decrease in wages—potentially owing to the devaluation of women's work [216]. Women often take on a disproportionate amount of unpaid care work compared with men, a factor that has historically been overlooked when examining gender pay gaps [217]. The unpaid care work and time devoted to family-oriented goals is one of the factors cited as contributing to the difficulty of applying established career-trajectory models to female workers [218]. Indeed, these differences may manifest—at least partially—as an increased proportion of women working part-time, with women working part-time at roughly twice the rate of men [219]. Mothers are slightly more likely than fathers to report that they felt like they needed to reduce work hours or turn down a promotion due to family reasons [214]. In academia, there appears to be a parenthood penalty for mothers more so than fathers in terms of the productivity effects of having children [220]. The effects of parenthood on employment have been especially stark during the COVID-19 pandemic, with childcare issues negatively impacting women's professional lives at greater rates than men [221]. Additional factors are only beginning to be understood, such as the preference for shorter commutes by women, its relation to family responsibilities, and ultimately wages [222].

Occupation selection and disparities in specific occupations along gender lines drives some of the overall gender pay gap [215]. In science, technology, engineering, and mathematics (STEM) there is a notable gender imbalance—with women having higher representation in biological and chemical sciences and men tending towards

mathematics and computational occupations. The imbalance is nuanced within fields, for instance in computer science women are more prevalent in human computer interaction and men are overrepresented in robotics [223]. Some of this gap may be driven by gender stereotypes that are instilled in children from a young age, with research finding that men and boys are more likely to view themselves as talented [224].

Hiring practices may also negatively impact mothers at higher rates than fathers. Correll *et al.* [225] show that changing parental status on application materials reduces the wage and perceived competence scores attributed to mothers, whereas men may experience a slight benefit from being parents. In the EU labor market fatherhood has been found to have a wage premium whereas there are varied impacts of motherhood—but it generally decrease women's compensation [226]. Recent research from the Swedish labor market suggests that male applicants to jobs in female-dominated occupations have lower response rates than their female counterparts (with balanced and male-dominated occupations having near parity in response rates) [227]. One study of resume rankings on the job sites Indeed, Monster, and CareerBuilder found modest but statistically significant biases that benefited men both at the group and individual level (for a subset of titles and labor markets explored) [228].

Job postings and resumes are perhaps the primary artifacts that companies and job seekers interact with when filling roles. Job postings have been shown to relate to the average wages of professionals and firm productivity [229]. Other work has used large-language models trained on job postings to explain some of the variance in offered wages [230].

There are relatively fewer large-scale studies of resume content and how it relates to worker compensation or hiring decisions. Small scale studies have suggested there

is an interaction between applicant, recruiter gender, and how resumes are perceived, with female recruiters being more likely to rank male applicants as having more work experience [231]. Another small-scale study from 1975 found that female applicants are perceived as less competent than their male counterparts [232]. Similarly sized studies found inconclusive results for a set of resume reviewers—with varied inferences made based on hobbies, professional formatting, and reference writers among the reviewer panel [233].

When conducting research in the area of AI fairness, it is important to define our working definitions of bias and harm. For the current piece, we are primarily concerned with quantifying the gender biases present in resume data. More specifically, in our case gender bias takes the form of *meaningful* differences in the textual data associated with female and male workers—where *meaningful* is in turn informed by the biases' relation to known harms. The primary harm we are concerned with is the gender pay gap. Female workers earning less than their male counterparts presents tangible harms related to economic hardships and is indicative of issues related to career advancement. Secondary (and related) to the wage gap, is the gender representation gap in a given field. Aggregate gender representation in a given field does not present the same level of specific harm as wage. However research suggests gender representation imbalances can be a precursor to wage disparities [216]. Additionally, there are concerns about feedback loops in fields being viewed as stereotypically favoring female or male workers, subsequent discrimination, and trainees having access to peer mentorship. Finally, seminal studies on the gender bias present in language models use the proportion of female to male workers in a field as reference point when demonstrating gender biases [44, 46, 208].

We build on prior studies studies in two primary areas: 1) we analyze real world resume data, examining the information that is used in practice when making hiring decisions; and 2) we include wage gap as a measurement of harm. To be clear, our work operates at the data set level—identifying biases present in the data that could manifest as biased AI pipelines (e.g., for candidate screening, job recommendation, etc.).

In the current study, we quantify the language differences between female and male workers using data from their resumes. We relate these differences to broader labor market conditions, such as the gender pay gap and gender representation in specific jobs. To do this we combine a large data set of resumes with demographic data and labor economics data. Combining these data sets allows us to conduct regression analyses to determine the extent to which language differences between female and male workers describe the gender pay gap in the US labor market. In Sec. 4.3 we introduce our data sets and describe our analytical methods. In Sec. 4.4 we present our results, describing differences in female and male language usage and their association with the gender pay gap. Finally, in Sec. 4.5 we summarize our findings and suggest directions for future research.

4.3 Methods

4.3.1 Data

We combine data from a large collection of resumes from the US labor market with population level statistics from the US Bureau of Labor Statistics (BLS) to conduct

the current study. In addition, we use data from the Occupational Information Network (O*NET) to provide work place activities associated with specific occupations.

Throughout the study, we use the BLS 2018 Standard Occupation Classification (SOC) system as a consistent taxonomy of job titles [234]. The SOC system provides a widely adopted standard for describing and grouping occupations that enables us to aggregate and compare results in a principled fashion. Within the SOC there are major, minor, broad, and detailed occupations—with detailed occupations representing the lowest level of the taxonomy. We use the example titles included with the SOC that correspond to each detailed position to generate word-embeddings for the positions. We then use the detailed occupation word-embeddings for matching worker-provided job titles to one of the 867 detailed occupations for further analysis. See Fig. 4.1 for a visual summary of the example title embedding space and list of major occupation categories.

The Occupational Information Network (O*NET) provides detailed work activities (DWAs) that consist of sentence-level descriptions of typical activities performed within given job families [235]. Using a crosswalk between BLS SOCs and O*NET job classes we create sets of DWAs associated with each SOC—these collections of DWAs are used to created word-embeddings to generate DWA clusters representative of each occupation.

Resume data is provided by FutureFit AI. The FutureFit AI data includes machine readable information on over 200 million resumes from the US labor market. In the resume dataset we have information on worker gender, education, location, and self-reported positions held over the course of their career. For each position on a resume

there are fields for dates of tenure, job title, location, company, and a free text description of the position.

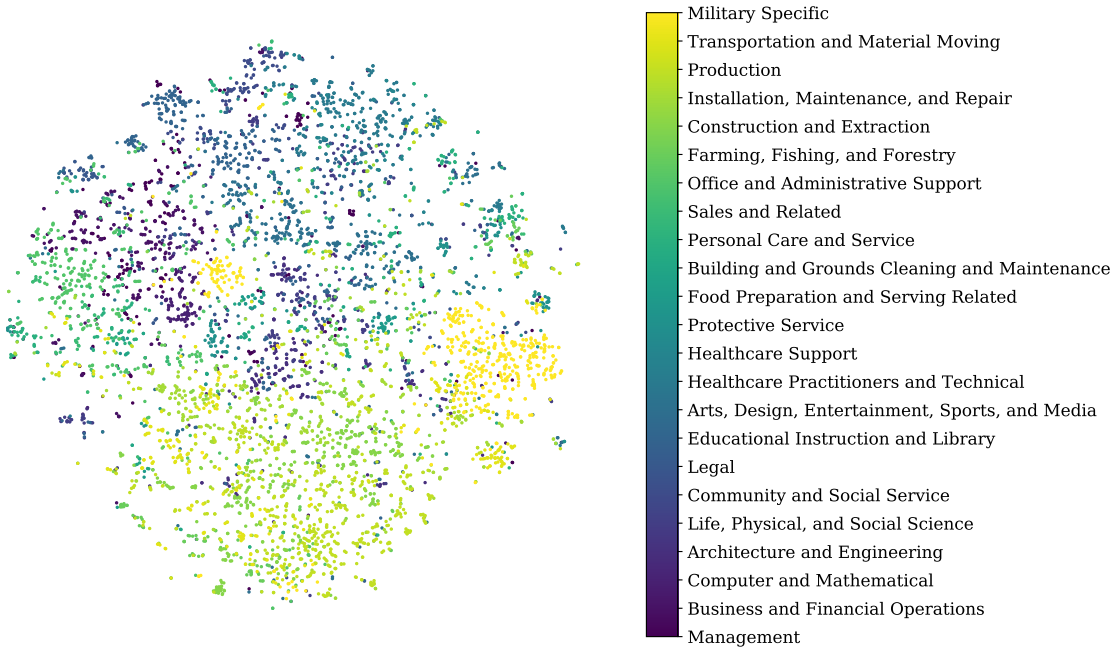


Figure 4.1: ***t-distributed stochastic neighbor embedding (tSNE) of example job titles using sentence BERT (SBERT)****. The embedding visualization provides a general indication of semantic similarity of job titles in the SBERT semantic space. Points represent an individual example job title provided by the US Bureau of Labor Statistics Standard Occupational Classification (SOC) system. Points are colored based on their membership in one of 23 major occupation categories.*

4.3.2 Language distribution divergence

We use rank-turbulence divergence (RTD) [3] and Jensen-Shannon divergence (JSD) to quantify the differences in language distributions between female and male workers' resumes. These divergence values allow us to highlight some the most salient differences in language usage by female and male applicants based purely on empirical observations of language use frequency (see Fig. 4.2 for an example). Divergence

values in general, namely RTD and JSD in our case, are helpful both for identifying specific words or phrases that have meaningfully different usage between language distributions, as well as summarizing the overall divergence between two distributions.

The rank-turbulence divergence between two sets, Ω_1 and Ω_2, is calculated as follows,

$$\begin{aligned} D_{\alpha}^{R}(\Omega_1||\Omega_2) &= \sum \delta D_{\alpha,\tau}^{R} \\ &= \frac{\alpha+1}{\alpha} \sum_{\tau} \left| \frac{1}{r_{\tau,1}^{\alpha}} - \frac{1}{r_{\tau,2}^{\alpha}} \right|^{1/(\alpha+1)} , \end{aligned}$$

where $r_{\tau,s}$ is the rank of element τ (n-grams in our case) in system s and α is a tunable parameter that affects the impact of starting and ending ranks.

The Jensen-Shannon divergence between two sets, Ω_1 and Ω_2, is calculated as follows,

$$\mathrm{JSD}(\Omega_1||\Omega_2) = \frac{1}{2} D(P_1||P_{\mathrm{M}}) + \frac{1}{2} D(P_2||P_{\mathrm{M}}) \tag{4.1}$$

Where, $D(P_x||P_y)$ is the Kullback-Leibler divergence between probability distributions P_x and P_y, $P_{\mathrm{M}} = \frac{1}{2}(P_1 + P_2)$, and P_i is the probability distribution for set Ω_i.

To generate the language distributions divergences we start by counting the occurrences of 1-grams (words, basically) in resumes delimited by gender and detailed occupations. We then calculate the JSD and RTD between female and male language distributions. For the analysis of differences between genders we mainly compare language within detailed occupations to control for differences in occupation specific language usages and associated gender imbalance.

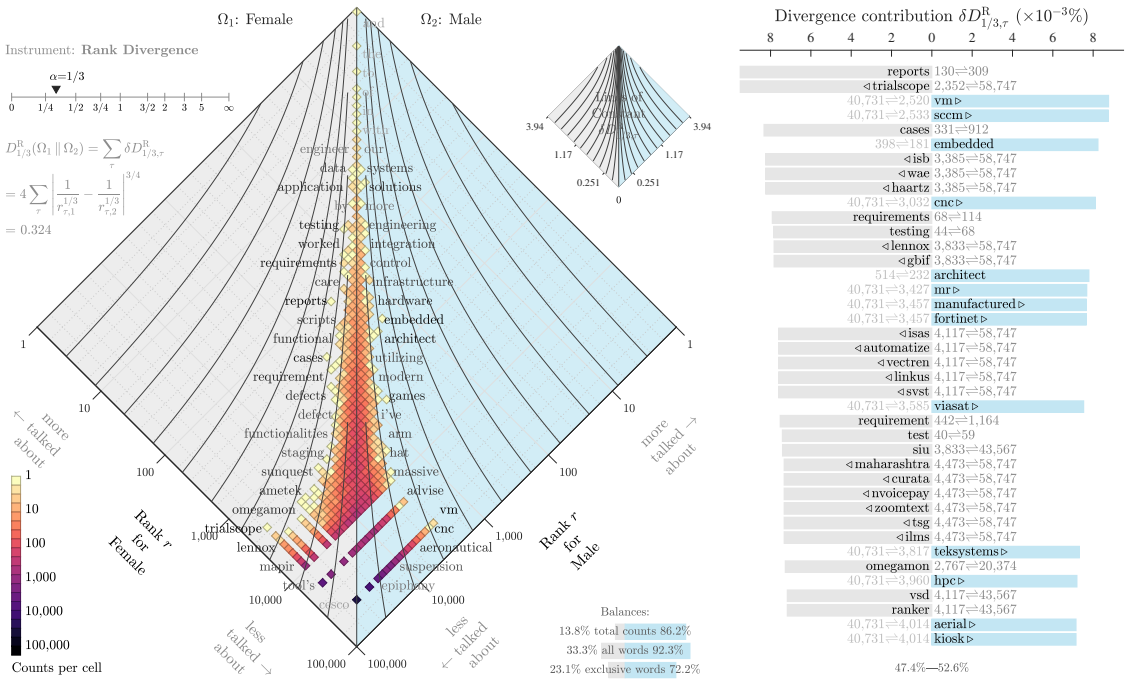


Figure 4.2: ***Allotaxonograph [3] for female and male software developers (detailed occupation).*** *The central diamond shaped plot shows a rank-rank histogram for 1-grams appearing in each gender's resume language distributions. The horizontal bar chart on the right shows the individual contribution of each 1-gram to the overall rank-turbulence divergence value ($D^R_{1/3}$). The triangles to the right or left of* 1*-grams indicate that the term is unique to that distribution. The 3 bars under "Balances" represent the total volume of 1-gram occurring in each distribution, the percentage of all unique words we saw in each distribution, and the percentage of words that we saw in a distribution that were unique to that distribution.*

4.3.3 Skills embeddings

We match text descriptions of positions from resumes with DWAs from O*NET using sentence BERT (SBERT) [236]. SBERT is a BERT-based language model optimized for semantic similarity tasks. SBERT allows us to establish a link between latent skills present in resumes and canonical skillsets from O*NET. In this we have SBERT embeddings of worker resumes, $\overrightarrow{r}$, and O*NET detailed workplace activities (DWAs), $\overrightarrow{a}$, for a given SOC. More specifically,

$$\overrightarrow{a_i} = \text{mean}_{i \in soc}(\overrightarrow{\text{activities}_i})$$

To match resumes with SOCs using their associated DWAs, we take the maximum cosine similarity score between a resume embedding $\overrightarrow{r}$ and all pairwise comparisons with activities embeddings $\overrightarrow{a}$. Assigning resume r_j to a detailed occupation with the following:

$$SOC_j = \underset{i}{\text{argmax}}(\cos(\overrightarrow{r_j}, \overrightarrow{a_i})) \qquad , \forall i$$

The detailed occupation with a DWA cluster that is most similar to the resume embedding is assigned to the resume. We compare these results with the title-matched detailed occupations assigned from Sec. 4.6.

4.3.4 Word-Embedding Association Test

We use the methodology from Caliskan *et al.* (2017) [44] to measure the association of 1-grams from resumes with the concept of gender. At its core, the Word-Embedding

Association Test (WEAT) relies on cosine similarity scores between attribute words and target words. Examples of attribute categories from Caliskan *et al.* include pleasant vs. unpleasant, temporary vs. permanent, and male vs. female. Examples of target categories include instruments vs. weapons, math vs. arts, and young vs. old people's names. In our case, the attribute words are male vs. female and the target words are the top 50 words contributing to RTD from the female and male language distributions (100 words total) for a detailed occupation.

The association is measured by comparing distributions for women and men with the gender attribute terms and looking at the standardized difference between the similarity scores. More specifically, the test statistic for each target word w is defined by

$$s(w, A, B) = \frac{\text{mean}_{a \in A} \cos(\overrightarrow{w}, \overrightarrow{a}) - \text{mean}_{b \in B} \cos(\overrightarrow{w}, \overrightarrow{b})}{\text{std\_dev}_{x \in A \cup B} \cos(\overrightarrow{w}, \overrightarrow{x})}$$

The difference in $s(w, A, B)$ for all target words associated with the female and male distributions provides the final bias measure. This is defined by

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in y} s(y, A, B)$$

Where X and Y are our target terms from female and male resumes. A and B are attribute terms corresponding to female and male genders.

We use the pretrained `word2vec-google-news-300` model from the original word2vec paper for all WEAT tests [13]. While the model is older and far from state-of-the-art, the main point is to gain a measure of gender association from a word embedding space, a task that is well understood with an older model such

as word2vec. Caliskan *et al.* (2017) find similar results when using the GloVe [7] word-embedding model.

4.3.5 Regression analysis

We conduct regression analyses to determine the impact of language distribution divergences between female and male workers on the wage gap for detailed occupations. For this analysis we fit a model predicting women's wages as a percentage of male wages, $G_f = W_f/W_m$, where W_F and W_m are the median weekly wages for women and men, respectively. We include observations of detailed occupations for the 12 years spanning from 2005 to 2017. The full model includes the following features:

- $\text{JSD}_{\text{gender}}$ and $\text{RTD}_{\text{gender}}$: Rank-turbulence divergence (RTD) Jensen-Shannon divergence (JSD) values for language distributions of female and male works in detailed occupations.
- $\text{JSD}_{\text{major}}$ and $\text{RTD}_{\text{major}}$: RTD and JSD values for language distributions of detailed occupations and the language distribution of major occupation (with the distribution of the given detailed occupation removed from the major distribution).
- w2v bias: Word-Embedding Association Test (WEAT) effect sizes.
- state quotient: state employment quotients for 50 states and 6 territories—defined as the percentage of a state's worker force employed in a detailed occupation divided by the national percentage of workers in that detailed occupation.

- Female emp.: Gender balance, GB, of female and male workers in the detailed occupation ($GB = N_f/N_m$).
- Major SOC FE: Fixed effects for SOC major occupation categories.
- Year FE: Year fixed effects.

4.4 Results

4.4.1 Resume topics

We fit top2vec topic models for a sample of resume position descriptions from each major occupation. In Fig. 4.3 we show the top 10 topics for the Computer and Mathematical Occupations major SOC. To determine the top topics and the gender composition of each topic, we assign positions from resumes to the nearest topic. After assigning positions (and associated workers) to each topic, we can calculate the gender balance of each topic.

In the case of the example topics in Fig. 4.3, we see an overall imbalance towards male workers—with roughly 26% of Computer and Math positions in our dataset belonging to female workers. In topic 10, we can see 1-grams that are associated with cybersecurity such as "vulnerability", "thread", and "intrusion". Topic 10 is also a male-dominated topic with only 13% of associated resumes belonging to female workers. On the other hand, topic 7 contains 1-grams related to clerical work, including words such as "entered", "filling", and "paperwork" The clerical topic has more female and male resumes associated with it, with 61% of this topic's workers being female.

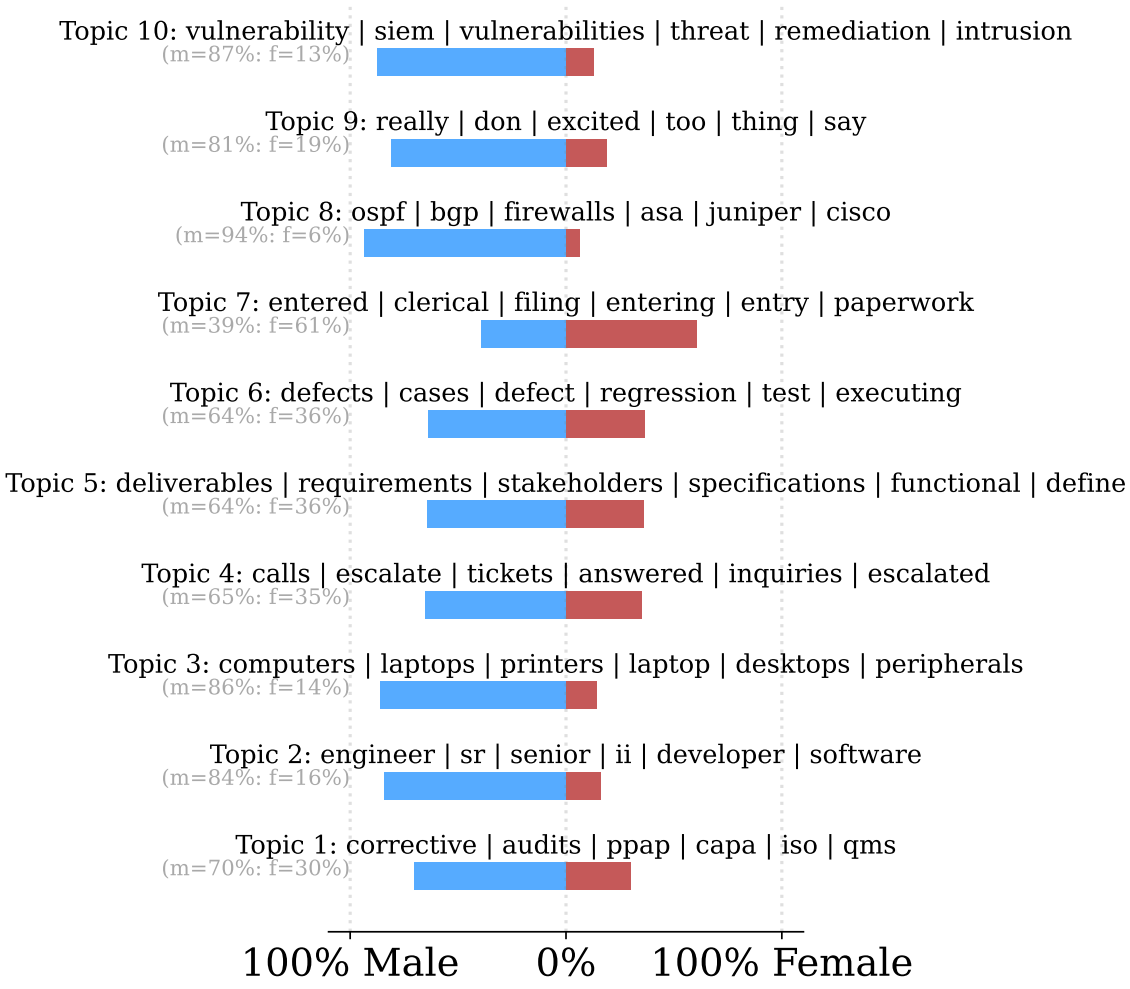


Figure 4.3: ***Topics for the mathematics and computer occupations major SOC category.*** *The bars represent the balance of women and men for each topic as determined by assigning resumes to their most similar topic. Associated words are the 6 most similar 1-grams in the joint 1-gram and topic embedding space produced by top2vec. Topics are ordered from most prevalent (topic 1) to least prevalent. See Fig. 4.11 for a visualization of the topic space.*

4.4.2 Similarity of job descriptions to canonical activities

We find notable differences in the semantic similarity of position descriptions on resumes and the corresponding set of detailed work place activities (DWAs) provided by

O*NET. We evaluate this difference by embedding job descriptions from resumes and DWAs using SBERT. The distribution of cosine similarity scores provides an indication of how close job descriptions from resumes are to O*NET DWAs in the semantic space—in theory, higher similarity scores represent worker-generated descriptions that are closer to expected descriptions.

Fig. 4.4 shows the rank distribution of true titles matched using the DWA approach for 6 major occupation classes. Kolmogorov-Smirnov tests show where there are significant deviations between female and male distributions. For instance, for computer and mathematical occupations males' true job titles are more important (lower rank value) than females' when using the position descriptions and DWA similarity method. Alternatively, for office and administration jobs, females' true position is more important than males' when using the DWA similarly approach.

4.4.3 Gender language divergence

Divergence results for language distributions corresponding to female and male workers within the same detailed occupations reveal some salient differences between genders. As an example we look at the software developer detailed occupation. Fig. 4.2 shows how specific words contribute to the divergence between the language distributions for female and male software developers. Overall, we notice an abundance of unique or somewhat rare words in the divergence contributions. Terms such "sccm", "gbif", and "linkus" refer to specific tools, organizations, or products. Indeed, there is a rather heavy tail of rare terms owing partially to the prevalence of specific tools and technologies in the field. We see more common words such as "requirements", "cases", and "tests" appearing more frequently in female resumes, which likely corresponds to

a higher prevalence of female developers in the sub-fields of quality assurance. On the other hand, we see the words "embedded" and "hardware" appear more frequently in the male distributing, likely owing to the higher prevalence of men in sub-specialities that deal more directly with computer hardware.

Moving to a different detailed occupation, Fig. 4.7 provides a counter example from the legal profession where we see less unique words driving the divergence. In this case, we see the female distribution contains relatively more occurrences of terms related to family law (e.g., "family", "domestic", "guardianship"). The distribution for male lawyers contains relatively more words related to intellectual property (e.g., "intellectual", "secrets") and personal injury (e.g., "accident", "injury").

4.4.4 Inter-occupation language divergence

To contextualize the divergence results presented throughout this piece, we calculate the pairwise language distribution divergence for all detailed SOCs (including workers from both genders). In Fig. 4.10 we present a tSNE visualization of these divergences but embedding them in a 2-dimensional space—in the figure the grouping of jobs by major SOC is apparent.

For context, using the RTD measure, two of the most similar detailed occupations are `Appraisers and Assessors of Real Estate` and `Real Estate Sales Agents` (RTD = 0.303). Notably, these occupations are from different major occupation groups, but due to their domain specific experience and skills the RTD measure is relatively small between them. On the other hand, two of the most different detailed occupations—as determined by RTD—are `Ophthalmologists, Except Pediatric`

and `Cutting, Punching, and Press Machine Setters, Operators, and Tenders, Metal and Plastic` (RTD=0.862).

4.4.5 Regression analysis

Differences in language usage by female and male workers have modest but statistically significant correlation with female wage share ($r = 0.29, p < 0.001$, see Fig. 4.5). On its own, Jensen-Shannon divergence describes roughly 11% of the variation in the female wage gap (See Table 4.1). The effect remains when controlling for women's employment share and the language differences of detailed occupations relative to the occupation's major occupation class, with this model accounting for roughly 22% of the variation in the wage gap. Our full model includes fixed effects for year and major occupation as well as state employment quotients. Adding in the language divergence feature for gender improves classification for the full model, as evidenced by the increased $R^2 = 0.546$ value. Further, the coefficients for the gender language distribution are relatively stable and statistically significant at the $p < 0.001$ level for for all model types. See Fig. 4.6 for a comparison of model root mean squared errors.

In an effort to control for gendered language, we include a WEAT variable (`w2v bias`) that describes the bias of 1-grams for the distributions corresponding to each observation. On its own, `w2v bias` accounts for roughly 6% of the variance in the wage gap, but after adding the variable to the full model we find the coefficient is pushed to effectively 0 and there is no change in variance explained. In contrast, when we fit a model to predict the gender employment gap, `w2v bias` describes roughly 18% of the variance by itself. In the employment gap model, the `w2v bias` coefficient

remains significant when included in the full model, and improves the R^2 value by roughly 0.03.

Qualitatively, the employment gap results are consistent with Caliskan *et al.* who find that there is a positive correlation between word-embedding gender bias of job titles and the gender distribution of a detailed occupation ($r = 0.90$, $p < 0.001$). In our case, we find that the word-embedding gender bias of the top divergence-driving words and the gender distribution of occupations is positively correlated as well, but with a smaller effect size ($r = -0.38$, $p < 0.001$, see Fig. 4.8).

4.5 Conclusion

In the current piece we describe aggregate level language differences between female and male workers' resumes. We go on to demonstrate an association between these differences and the gender pay gap. We show the gender composition of resume topics for a major category of jobs, shedding light on the signal present in the resume-text data set and its relation to gender (im)balances. We also present follow-on work from prior research that expands on the connection between gender bias present in semantic spaces with on-the-ground realities of gender distributions within occupations. Taken together these results demonstrate the value in leveraging text data to improve our understandings of workers' skills, experience, and other factors. We show that textual data contains signals that help describe real world harms such as the wage gap. Further, we see how one must be thoughtful about constructing text-based features—in our case wage gap and employment share required two different lenses to surface meaningful associations.

Dependent variable:								
Female wage percentage								
Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Female emp.	0.081*** (0.009)		0.074*** (0.009)	0.079*** (0.009)		0.057*** (0.011)	0.057*** (0.011)	0.059*** (0.011)
JSD_{gender}		0.185*** (0.021)	0.317*** (0.039)			0.221*** (0.044)	0.221*** (0.044)	
JSD_{major}			-0.493*** (0.103)			-0.212* (0.108)	-0.214* (0.109)	
RTD_{gender}				0.289*** (0.038)				0.185*** (0.052)
RTD_{major}				-0.094 (0.069)				0.152 (0.124)
w2v bias					-0.087*** (0.013)		-0.001 (0.014)	
State quotient						Yes	Yes	Yes
Year FE						Yes	Yes	Yes
Major SOC FE						Yes	Yes	Yes
R^2	0.100	0.114	0.230	0.189	0.063	0.556	0.556	0.543
Adj. R^2	0.099	0.112	0.227	0.185	0.061	0.490	0.489	0.475
$p < 0.1^*,\ p < 0.01^{**},\ p < 0.001^{***}$								

Table 4.1: ***Ordinary least squares models predicting the wage percentage for women.*** *Coefficient values are reported with standard errors in parentheses. State quotients, year fixed effects (FE) and Major SOC FE are reported as binary inclusion.*

The first indication of meaningful differences in language usage between genders is seen with the divergence measures. We show that the individual words contributing to the divergences are noteworthy and that viewing the words provides explainability of the divergence results. The next indication of gender disparities came with our analysis of matching canonical workplace activities for specific jobs with the text portion of position descriptions. The activity matching shows differences for female and male workers—suggesting that the idealized candidate for specific jobs may have gender bias, at least relative to empirically observed resume data. Topic modelling provided an indication of how resume text commonly clustered together along with the gender composition of these clusters. In an attempt to more meaningfully investigate biases, it is important examine the harms that may result [173]—in our case gender pay gap and gender representation within occupations are two examples of harm.

Our regression analysis demonstrates that job-specific language divergence measurements (JSD and RTD) detect salient differences between female and male workers' resumes with respect to the gender wage gap. These differences are likely not simply a function of gender employment share or other factors such as location, year, or major occupation. Indeed, the divergence measures are poor predictors of female employment percentage as seen in Table 4.2 and the correlation in Fig. 4.9 ($r = 0.01$). Further, the divergence measures appear to capture information that is not entirely described by measurements of gender bias in a word-embedding space. Taken together, these results suggest that there are latent characteristics of language distributions for female and male workers that help describe at least part of the gender pay gap.

The job-specific language divergence measurements are likely influenced by a plethora of factors. Broadly, geographic variation in language usage, temporal effects, and industry effects are all factors that likely influence the language characteristics we report here. We make an effort to control for these specific factors using fixed effects for time and major occupation category along with including a control for detailed occupation employment at the state-level. There are also factors related to specific firms and educational institutions that may affect the language features of resumes—something we do not control for in the present study.

We have presented some evidence that the distribution divergences are contributed to by sub-specialization (e.g., quality assurance vs. hardware engineers for software developers). Indeed, no two workers are exactly the same. Experience, qualifications, and skills, all contribute to workers that are best thought of as having high dimensional characteristics. Language is one manner by which to capture salient information that would be difficult to collect in tabular form—especially given current data repositories. Text-as-data, when used correctly, can help use create improved representations of workers for the purposes of labor economics, future-of-work research, worker re-skilling, talent screening, and other areas. But first we need to understand how linguistic features interact with other worker attributes and may contribute to real world harms. The gender wage gap is but one example of these harms—one for which data was available for the present study—but there other worker attributes that should be understood in relation to textual data (e.g., race, parental status, age, etc.). Using free text descriptions could unlock a vast amount of information that is increasingly important in today's modern workforce—especially as jobs, skills, and careers change more rapidly than ever. Accessing the promise of this data must come in tan-

dem with understanding how biases and inequalities are contained within—both for the purposes of better understanding and mitigating these issues. For instance, other research has emphasized the limitations of efforts to debias language models [212], and our work suggests that salient gender signals exist outside the dimensions of gender that would commonly be encoded in a general purpose language model.

Thinking broadly about our results, we believe there is some logic behind the finding that the representation gap is associated with gender bias in the word-embedding space, while the wage gap would be associated with job-specific language differences. For one, the gender bias present in language models is likely owing to societal associations of genders and occupations based on the demographics of workers (this point can lead to feedback loops with subconscious selection for stereotypical workers). On the other hand, intra-occupation wage gaps are less discussed and likely less central to societal perceptions of jobs and worker characteristics. Instead, wage gaps within occupations are more likely related to factors that are also encoded in resumes—skills, experience, certifications, and so on.

We were limited to examining binary cases of gender due to the constraints on the available datasets (both the resumes and BLS data are coded with female and male genders). Future research should be conducted with non-binary cases of gender. Furthermore, we were not able to evaluate the accuracy of the gender variable in our resume data set—a potential issue given some gender classifiers have been shown to vary in accuracy along racial and ethnic lines [237]. Our resume dataset does not purport to be a totally representative sample of workers from the US labor market, and the potential for sampling bias influencing our results cannot be ruled out. We are also unable to comment on the direct causes of the wage gap based on our ana-

lytical framework—we are merely able to present notable associations. Future work could investigate how modifying language in resumes affects job screening and hiring processes—putting our theories to the test in more real world settings.

4.6 Supplementary Information (SI)

Job title classification

Using sentence BERT (SBERT) we create job title word-embeddings, $\overrightarrow{t}$, for all titles appearing in the FutureFit AI resume dataset. We apply a similar procedure to the BLS SOC example titles, starting by creating a word-embedding for each title (BLS provides 4-10 example titles for each detailed occupation). Next, we take the average embedding of the example job titles for each detailed SOC—this mean vector will represent the detailed SOC category for the resume title classification task. The collection of these mean vectors for occupations, $\overrightarrow{o}$, forms the set of all occupation mean vectors, O. Classification is performed by searching for the most similar SOC embeddings for each resume position title embedding, $\overrightarrow{t}$ in the resume title set T. Using cosine similarity, we calculate the pairwise cosine similarity between each title and all 867 detailed SOC embeddings. We take the `argmax` of the cosine similarly vector for each resume title, and assign the corresponding SOC job title as a candidate match,

$$\underset{i}{\operatorname{argmax}} \cos(\overrightarrow{t}, \overrightarrow{o}_i) \quad , \forall i$$

Finally, we threshold by calculating the empirical cumulative distribution function for all pairwise matches (i.e., not just the 'best' or most similar match) and retaining the candidate matches that are greater than 99.9% of all cosine similarity scores.

Dependent variable:								
Female employment percentage								
Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Female sal.	1.237*** (0.155)		1.314*** (0.157)	1.309*** (0.160)		0.765*** (0.154)	0.702*** (0.145)	0.742*** (0.153)
JSD_{gender}		0.106 (0.068)	-0.104 (0.186)			-0.496** (0.185)	-0.453** (0.176)	
JSD_{major}			-0.113 (0.534)			-0.209 (0.528)	-0.575 (0.508)	
RTD_{gender}				-0.192 (0.142)				-1.182*** (0.167)
RTD_{major}				-0.631* (0.334)				-0.314 (0.445)
w2v bias					-0.589*** (0.048)		-0.318*** (0.048)	
State quotient						Yes	Yes	Yes
Year FE						Yes	Yes	Yes
Major SOC FE						Yes	Yes	Yes
R^2	0.100	0.002	0.104	0.115	0.189	0.608	0.639	0.625
Adj. R^2	0.099	0.001	0.100	0.111	0.188	0.550	0.585	0.569
$p < 0.1^*,\ p < 0.01^{**},\ p < 0.001^{***}$								

Table 4.2: ***Ordinary least squares models predicting the employment percentage for women.*** *Coefficient values are reported with standard errors in parentheses. State quotients, year fixed effects (FE) and Major SOC fixed effects are reported as binary inclusion. Female salary (sal.) is the dependent variable from Table 4.1.*

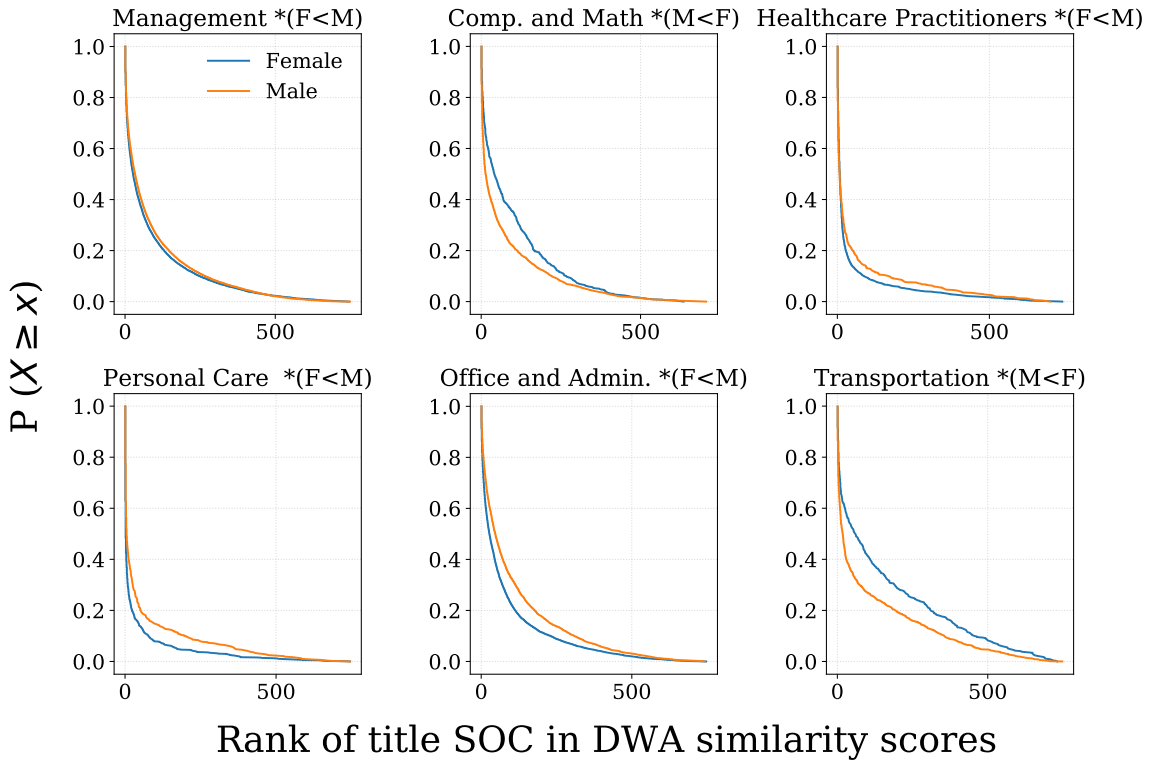


Figure 4.4: ***Rank distributions for job titles when inferred from resume descriptions matched against detailed workplace activities (DWAs).*** *Position descriptions from resumes are embedded using SBERT along with detailed workplace activities pertaining to each detailed SOC (provided by BLS). Rank values are for the cosine-similarity scores between DWA and resume embeddings. In general, a distribution being shifted to the right indicates that DWAs are less similar to the resumes belonging to that gender. Transportation has a larger difference in the female and male rank distribution than compared with Management. Stars indicate where difference in distribution is significant as detected by a one-sided Kolmogorov-Smirnov test.*

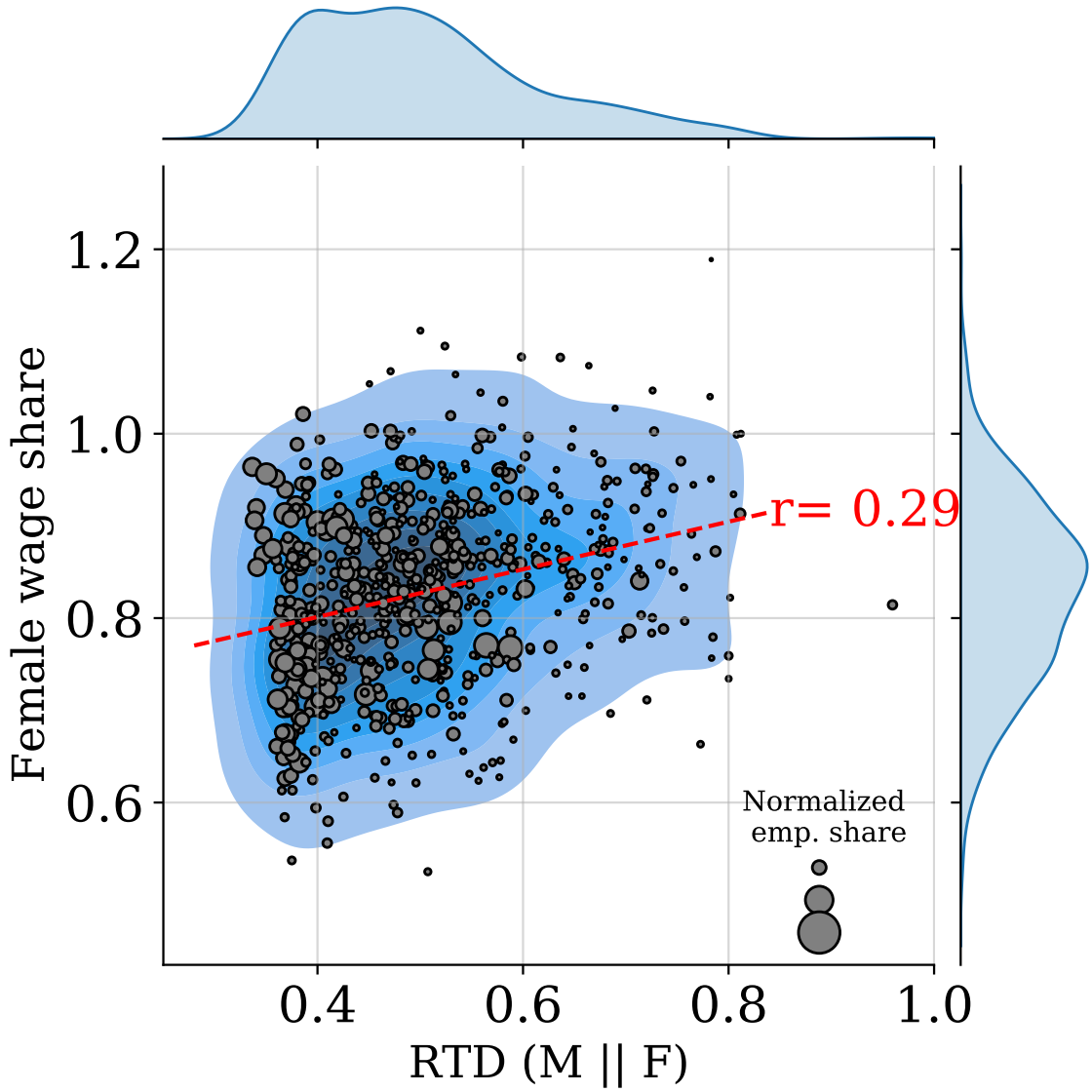


Figure 4.5: ***Pearson correlation between RTD and female wage share for detailed occupations from 2005 to 2017****. There is a positive correlation between women's earnings and the language divergence between women and men for a given detailed occupation in the data set.*

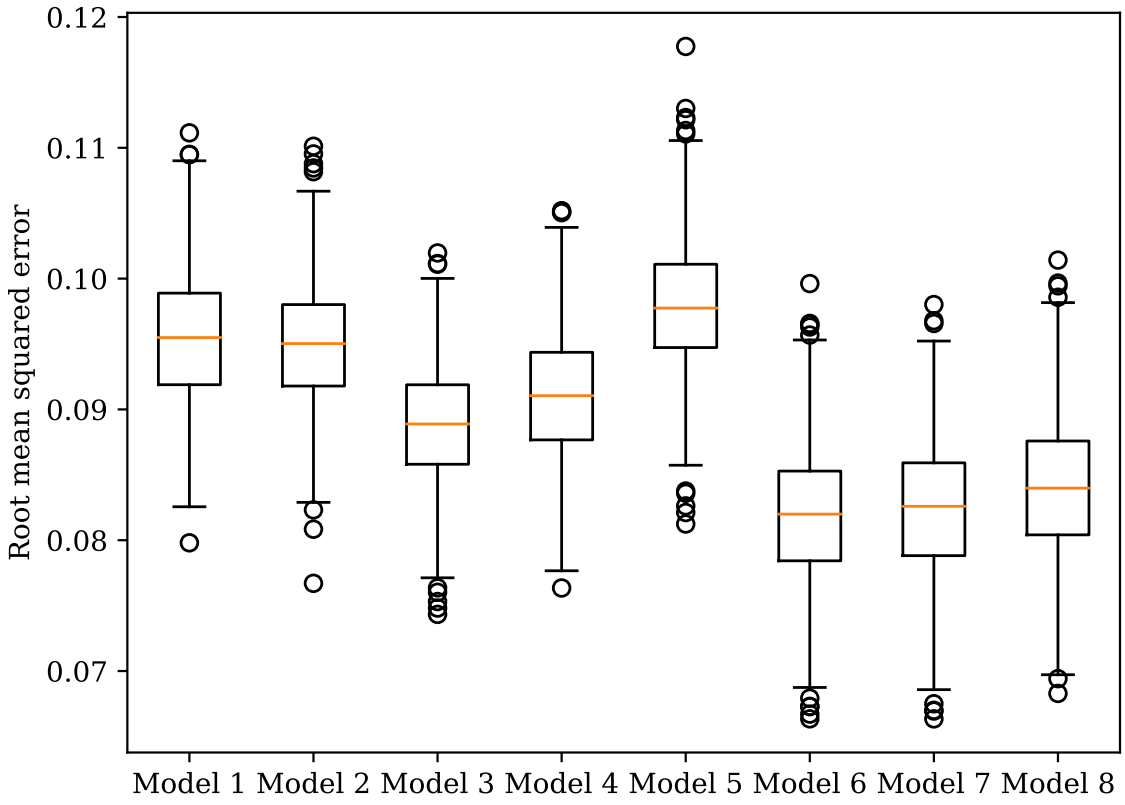


Figure 4.6: ***Root mean squared error for models predicting women's wage share.*** *See Table. 4.1 for model specifications and R^2 values.*

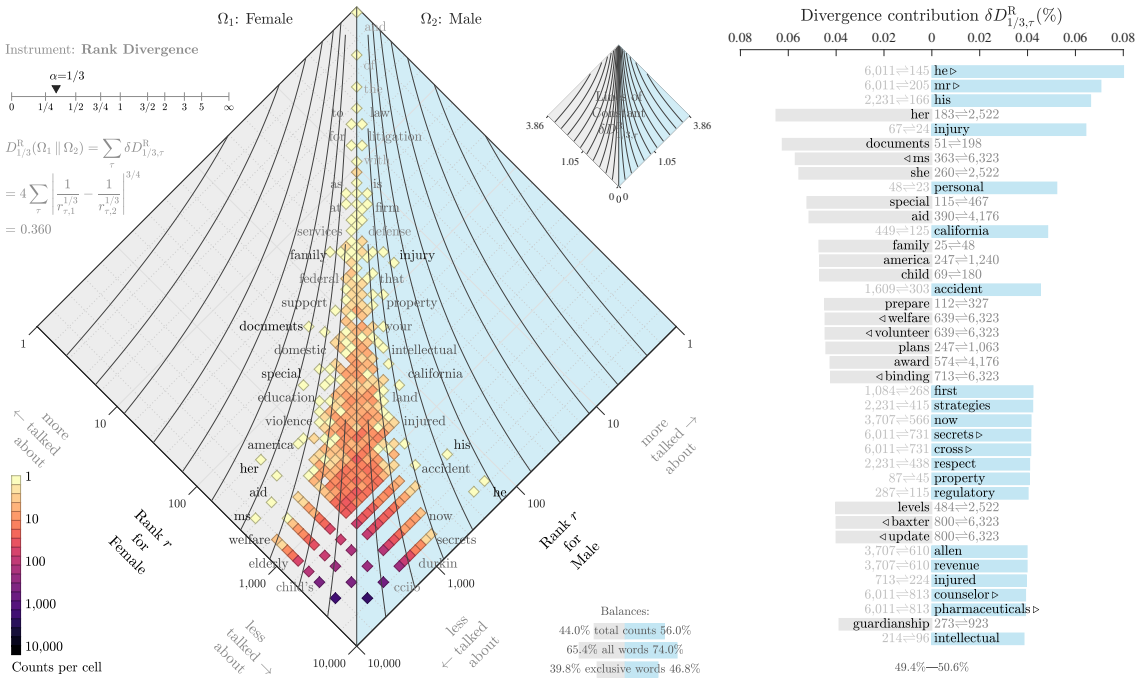


Figure 4.7: ***Alltotaxonograph [3] for the 1-gram distributions of resumes for female and male lawyers (detailed occupation).***

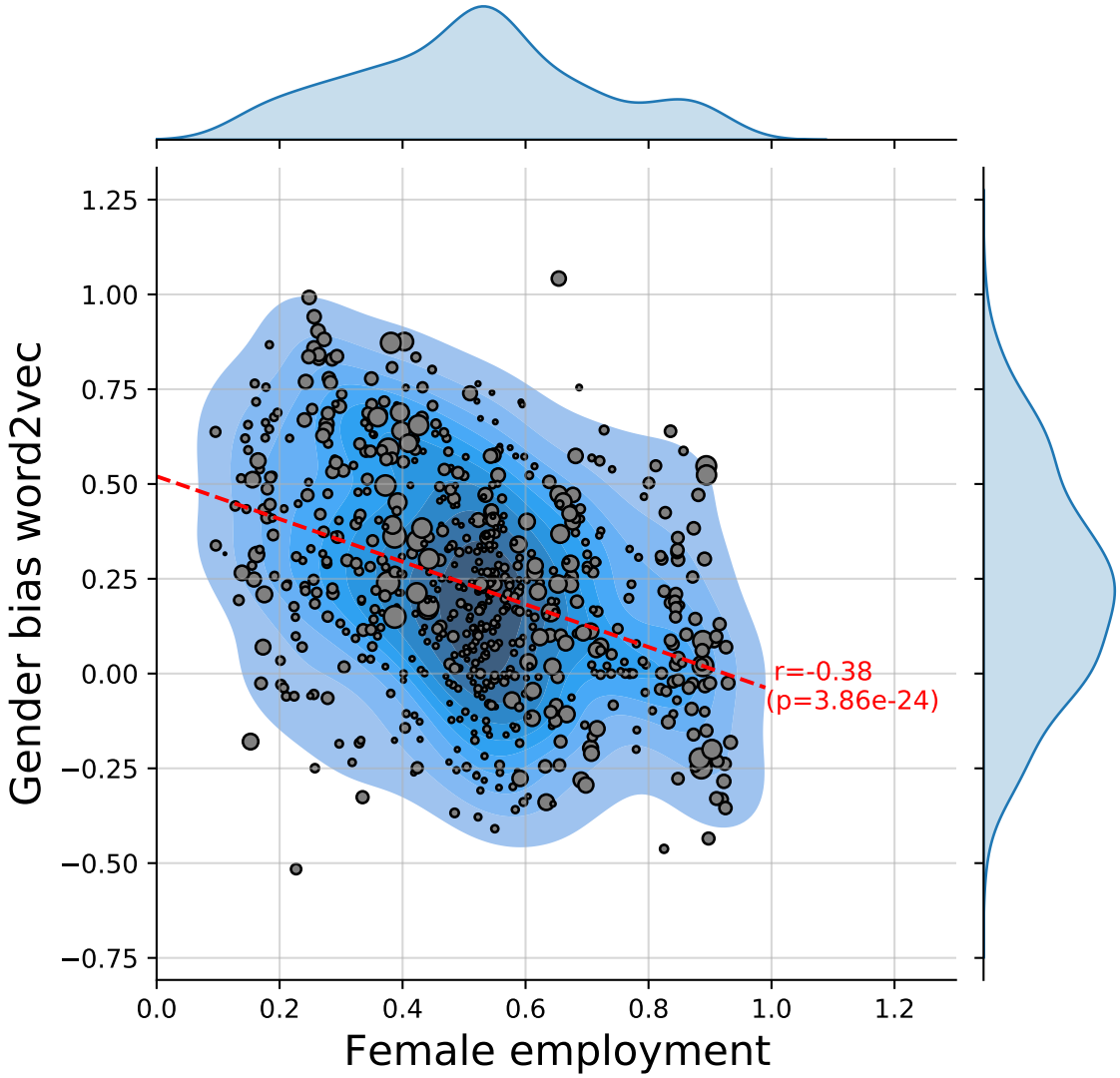


Figure 4.8: ***Bias scores for word2vec-derived gender bias effect compared with proportion of workers who are female for detailed occupations.***

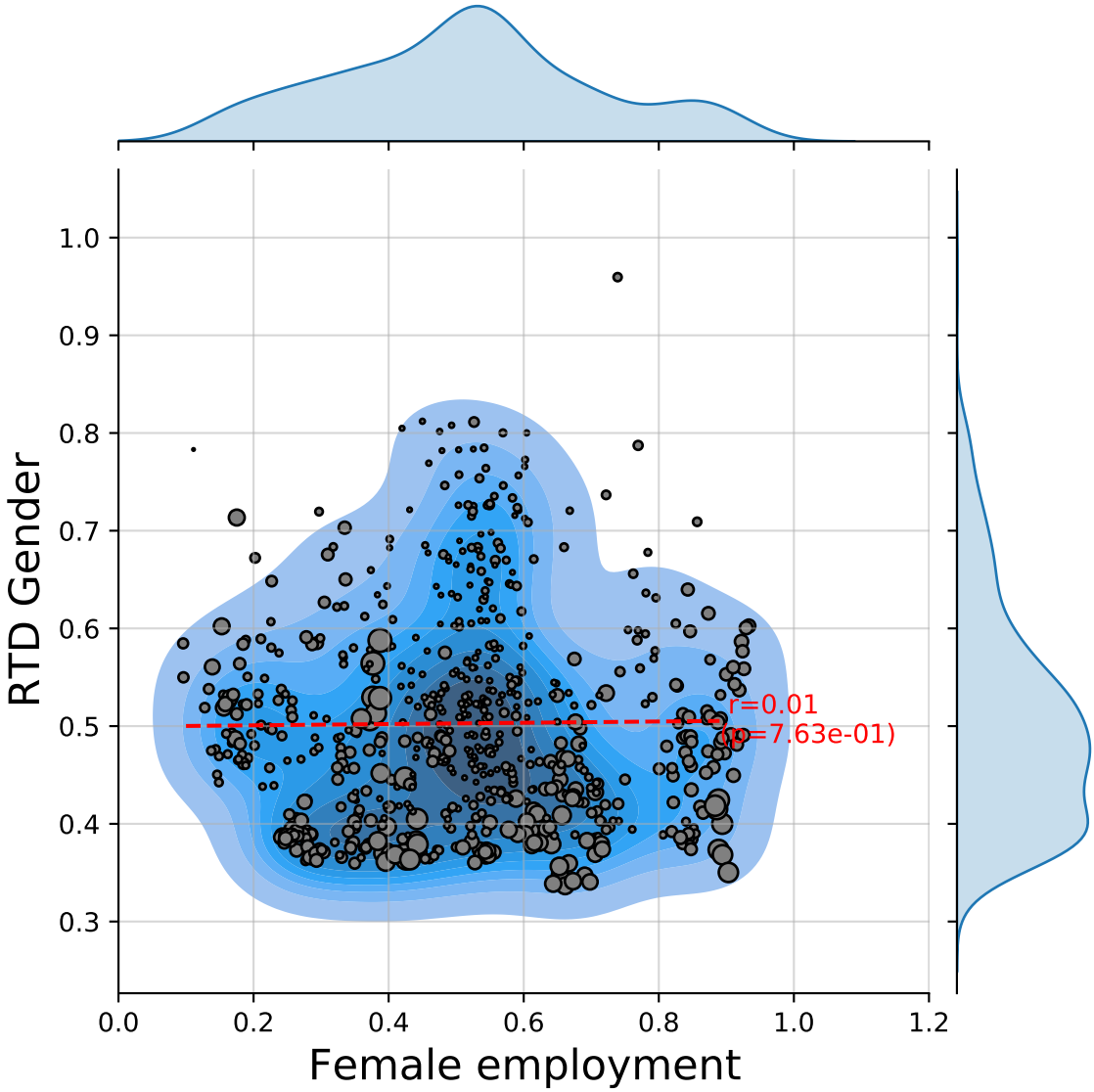


Figure 4.9: ***Rank-turbulence divergence over female employment share***

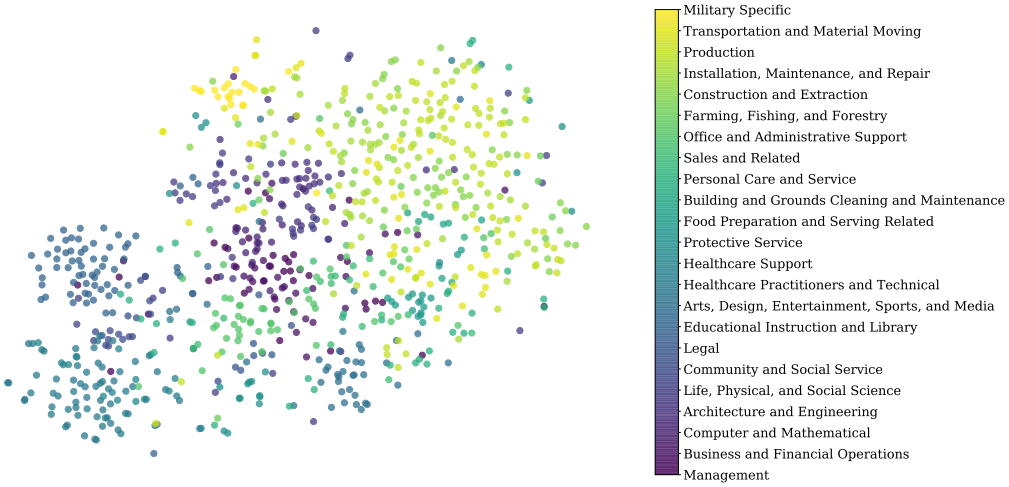


Figure 4.10: ***tSNE embedding of language distribution rank-turbulence divergences.*** *Each point corresponds to a detailed occupation and is colored by its corresponding major occupation. The location in the visualization is determined by an embedding created using the RTD divergence values as a pre-computed distance.*

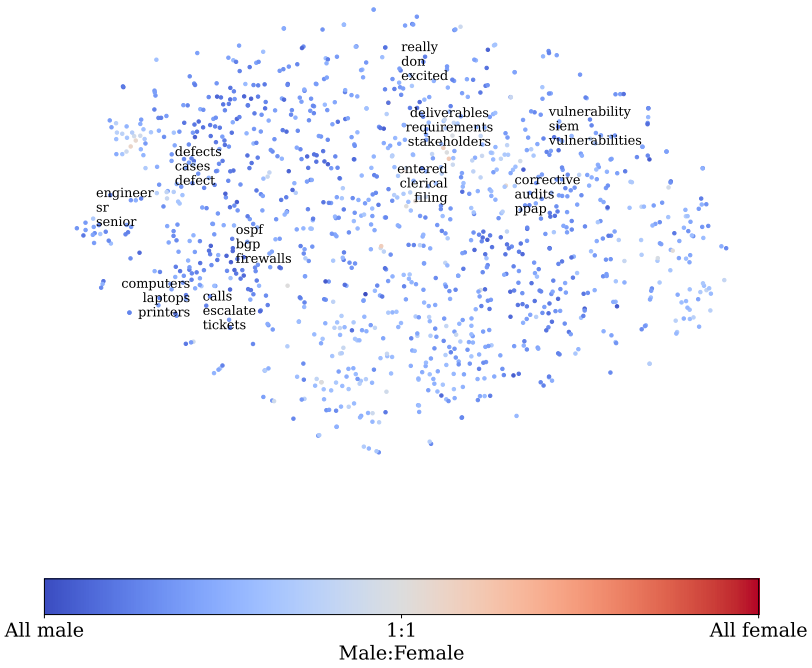


Figure 4.11: ***tSNE embedding of top2vec topics for resumes in the Computer and Mathematical Occupations major SOC.***

Chapter 5

Libraries

5.1 Overview

This section provides an overview of the libraries behind the research projects presented in this dissertation. Scientific computing is a central component of computational social science, but the software is not always discussed prominently.

5.2 List of libraries

- `twitter-utils`: a utility package for working with tweet JSON objects, managing flatfiles at scale, and scheduling jobs on a high performance compute cluster.

- `potusometer`: an API for interfacing with reply threads stored in MongoDB. The Ratiometer project (Chpt. 2) required the collection and indexing of hundreds of millions of tweets. Storing these tweets in MongoDB allowed for rapid

iteration of queries and answering more complicated research questions. This library abstracted much of the convoluted MongoDB queries away from the end user.

- `doctors-notes`: a library for augmenting textual documents based on complex characteristics of the document meta-data and the documents themselves. As part of the interpretable bias mitigation project (Chpt. 3) this library allowed for text documents to be readily augmented based on language distribution divergence measures. In addition, the library provides a facility for training a high number of language models to test the efficacy of the augmentation procedure.

- `soc-classification`: a library for classifying free text based on semantic similarity to the text of target classes. The library was primarily used for classifying job titles and detailed workplace activities in the resume bias project (Chpt. 4). The primary functionality is provided by the matching engine based on the sentence BERT (SBERT) [236] language model. In addition, the library provides functionality for locality-sensitive hashing for document deduplication.

Chapter 6

Conclusion

Here we present a collection of three works broadly concerned with improving our understanding of the representativeness of data that describes sociotechnical systems. The improved representation can be additional context of social media posts for a class of ultra-famous users, as in Chpt. 2. Another improvement can come with better understanding the biases present in the data and AI pipelines as with Chpts. 3 and 4. Chpt. 3 explores differences in language usage in electronic health records corresponding to female and male patients, and presents a methodology for reducing gender signals in the data. Chpt. 4 investigates how differences in the language distributions of resumes for female and male workers may be associated with the gender pay gap—in the process highlighting the subtleties of sub-specialities in many fields.

The motivations for understanding the dynamics of response activities in Chpt. 2 are multifaceted. For one, the ratio of these activities has become a cultural phenomenon on Twitter—more replies or quote tweets than likes or retweets can be seen as indicative of a controversial post. Second, the types of users we examined qualify

as ultra-famous users [123] on Twitter. Obama and Trump are examples of users who are talked about and engaged with at a rate that few others can match. With Twitter being a direct line to the public, it is important to understand how users respond (and spread) presidential communications on the platform. Finally, understanding response activities provides helpful context to data on social media. With this work it is possible to weight posts by the absolute value of responses, the ratio of these responses, or some function of response volume over time. We hope that a better understanding of the ratio leads to an additional filter that can be applied to lexical instruments operating on social media data.

More broadly, response activity dynamics may be helpful when making content moderation decisions and understanding the impact of recommendation algorithms on social media platforms. Future work in this area should address broader classes of users as defined by their fame on that platform. Further, it is worth pursuing analogous studies of response activities from other platforms and ideally developing a unified framework for cross-platform comparison.

We move to examining health records and large-language models in Chpt. 3, motivated by the goal of understanding the gendered dimensions of clinical texts and its interaction with language modelling pipelines. More generally, we are motivated to investigate demographic signals in text and the extent to which these signals could be reduced. We show that a divergence based measure can surface meaningful differences in language usage for female and male patients. We also find that using divergence results we can obscure the gender signal in patient notes—retaining classification performance on medically relevant tasks but preventing the identification of patient gender. Finally, we highlight the differences between gender bias as detected in the

word-embedding space with the bias detected on the data set using the divergence measure. The last point highlights the issue of tailoring our instruments to specific domains and tasks.

Going forward, studies could investigate how our data augmentation procedure compares (and perhaps interacts) with other debiasing techniques (e.g., adversarial debiasing [238]). Moving beyond binary definitions of gender and other demographic variables is another important next step for this research—a step that will require a modified framework and access to data sets with additional features. Looking towards fields other than medicine would help explore the generalizability of this framework. Given that interpretability was a central driver of the methodology, going forward we would like to evaluate how various stakeholders perceive and react to different bias mitigation approaches.

Our final study (Chpt. 4) investigates the connection between the position descriptions on resumes and the gender wage gap. The work is novel for making a large-scale assessment of the free-text portion of resumes and connecting the textual data to wage gaps and gender representation by field. Previous work in NLP had connected gender imbalances in fields with semantic biases—but to the best of our knowledge, we are the first to relate gender biases in the labor market with the actual text content of workers' resumes at scale. We find that the job-specific language differences are associated with gender wage gap, while semantic differences in the language used on resumes is associated with gender representation within occupations. One broader takeaway from this result is the value of more nuanced representations of workers in this data. The standard taxonomies of occupations fail to fully capture the details of sub-occupation specialization. Workers are high dimensional, having attributes like

skills and experience that are difficult to neatly quantify in tabular formats. Here we see the value of text-as-data, and the necessity of evaluating it critically to discover biases and improve our understanding of central processes in a sociotechnical system.

Similar to work in Chpt. 3, the framing of gender bias in Chpt. 4 fails to move beyond binary definitions of gender. A next step for the research is to investigate how to represent and study other framings of gender. Furthermore, future work should incorporate additional demographic variables and consider how to assess bias through an intersectional lens that factors in the interaction between these variables [239, 240]. It is worth considering how we could evaluate labor markets beyond the US. Moving beyond the US labor market may be difficult owing to available data and varying norms surrounding hiring. All of the studies discussed above were developed primarily with data from the US. While they offer examples of how we can improve our understandings of sociotechnical systems, it is important to remember the bias introduced by the US-centric view. Making sociotechnical instruments adaptable and effective across cultural contexts is a worthwhile future direction for all of these projects.

Bibliography

[1] Twitter. What are Promoted Tweets?, 2019.

[2] P. S. Dodds, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, D. R. Dewhurst, T. J. Gray, M. R. Frank, A. J. Reagan, and C. M. Danforth. Allotaxonometry and rank-turbulence divergence: A universal instrument for comparing complex systems. *arXiv:2002.09770 [physics]*, February 2020. arXiv: 2002.09770.

[3] Peter Sheridan Dodds, Joshua R Minot, Michael V Arnold, Thayer Alshaabi, Jane Lydia Adams, David Rushing Dewhurst, Tyler J Gray, Morgan R Frank, Andrew J Reagan, and Christopher M Danforth. Allotaxonometry and rank-turbulence divergence: A universal instrument for comparing complex systems. *arXiv preprint arXiv:2002.09770*, 2020.

[4] David Lazer, Eszter Hargittai, Deen Freelon, Sandra Gonzalez-Bailon, Kevin Munger, Katherine Ognyanova, and Jason Radford. Meaningful measures of human society in the twenty-first century. *Nature*, 595(7866):189–196, 2021.

[5] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE intelligent systems*, 24(2):8–12, 2009.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*, 2013.

[7] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[8] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits

of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[9] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[10] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[11] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[12] Diane Coyle. Socializing data. *Daedalus*, 151(2):348–359, 2022.

[13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[14] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[15] Robert Wolfe and Aylin Caliskan. Low frequency names exhibit bias and overfitting in contextualizing language models. *arXiv preprint arXiv:2110.00672*, 2021.

[16] George Kingsley Zipf. *The psycho-biology of language: An introduction to dynamic philology*. Routledge, 2013.

[17] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[18] ML Menéndez, JA Pardo, L Pardo, and MC Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997.

[19] John R Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957.

[20] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[21] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[22] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[25] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[26] Prakhar Gupta and Martin Jaggi. Obtaining better static word embeddings using contextual embedding models. *arXiv preprint arXiv:2106.04302*, 2021.

[27] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[28] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[29] Martin Müller, Marcel Salathé, and Per E Kummervold. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*, 2020.

[30] Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385, 2021.

[31] Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, Patrick Hall, et al. Towards a standard for identifying and managing bias in artificial intelligence. 2022.

[32] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[33] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[34] James Bagrow and Yong-Yeol Ahn. Network cards: concise, readable summaries of network data. *arXiv preprint arXiv:2206.00026*, 2022.

[35] Negar Rostamzadeh, Diana Mincu, Subhrajit Roy, Andrew Smart, Lauren Wilcox, Mahima Pushkarna, Jessica Schrouff, Razvan Amironesei, Nyalleng Moorosi, and Katherine Heller. Healthsheet: Development of a transparency artifact for health datasets. *arXiv preprint arXiv:2202.13028*, 2022.

[36] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. Interactive model cards: A human-centered approach to model documentation. *arXiv preprint arXiv:2205.02894*, 2022.

[37] Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, et al. The ai index 2021 annual report. *arXiv preprint arXiv:2103.06312*, 2021.

[38] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[39] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

[40] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, volume 1, page 2, 2016.

[41] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[42] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *arXiv preprint arXiv:2002.12327*, 2020.

[43] Janis Klaise, Arnaud Van Looveren, Giovanni Vacanti, and Alexandru Coca. Alibi explain: Algorithms for explaining machine learning models. *Journal of Machine Learning Research*, 22(181):1–7, 2021.

[44] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[45] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

[46] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

[47] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*, 2018.

[48] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*, 2019.

[49] Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. Hurtful words: quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120, 2020.

[50] Thayer Alshaabi, Jane L Adams, Michael V Arnold, Joshua R Minot, David R Dewhurst, Andrew J Reagan, Christopher M Danforth, and Peter Sheridan Dodds. Storywrangler: A massive exploratorium for sociolinguistic, cultural, socioeconomic, and political timelines using twitter. *Science advances*, 7(29):eabe6534, 2021.

[51] Peter Sheridan Dodds, Eric M Clark, Suma Desu, Morgan R Frank, Andrew J Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M Kloumann, James P Bagrow, et al. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8):2389–2394, 2015.

[52] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.

[53] Hal Roberts, Rahul Bhargava, Linas Valiukas, Dennis Jen, Momin M Malik, Cindy Sherman Bishop, Emily B Ndulue, Aashka Dave, Justin Clark, Bruce Etling, et al. Media cloud: Massive open source collection of global news on the open web. *ICWSM*, 2021.

[54] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, et al. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182, 2011.

[55] Eitan Adam Pechenick, Christopher M Danforth, and Peter Sheridan Dodds. Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one*, 10(10):e0137041, 2015.

[56] Stefan Wojcik, Adam Hughes, and Emma Remy. 19% of U.S. adults on Twitter follow Trump.

[57] Zeynep Tufekci. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Eighth international AAAI conference on weblogs and social media*, 2014.

[58] Jürgen Pfeffer, Katja Mayer, and Fred Morstatter. Tampering with twitter's sample api. *EPJ Data Science*, 7(1):50, 2018.

[59] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[60] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[61] Stefan Wojcik and Adam Hughes. How Twitter users compare to the general public, April 2019.

[62] Stephen E. Frantzich. *Presidents and the Media: The Communicator in Chief*. Routledge, London, July 2018.

[63] Samuel Kernell. *Going Public: New Strategies of Presidential Leadership*. CQ Press, Washington, D.C, 3rd edition edition, August 1997.

[64] Twitter. Target – Twitter Developers, 2019.

[65] Twitter. Political Content, 2019.

[66] Mike Isaac. The ratio establishes itself on Twitter. *The New York Times*, February 2018.

[67] Luke O'Neil. How to know if you've sent a horrible tweet. *Esquire*, April 2017.

[68] Merriam-Webster. Words we're watching: 'ratioed'. https://www.merriam-webster.com/words-at-play/words-were-watching-ratio-ratioed-ratioing.

[69] Data for Progress. The ratio richter scale, 2019.

[70] Yascha Mounk. Only cowards don't get ratioed, Oct 2019.

[71] Gray Victoria. Is the twitter ratio getting out of hand?, Jun 2019.

[72] Oliver Roeder, Dhrumil Mehta, and Gus Wezerek. The Worst Tweeter In Politics Isn't Trump. *FiveThirtyEight*, October 2017.

[73] Saud Alashri, Srinivasa Srivatsav Kandala, Vikash Bajaj, Roopek Ravi, Kendra L. Smith, and Kevin C. Desouza. An analysis of sentiments on Facebook during the 2016 U.S. presidential election. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 795–802, August 2016.

[74] J. Ko, H. W. Kwon, H. S. Kim, K. Lee, and M. Y. Choi. Model for Twitter dynamics: Public attention and time series of tweeting. *Physica A: Statistical Mechanics and its Applications*, 404:142–149, June 2014.

[75] Ryota Kobayashi and Renaud Lambiotte. TiDeH: Time-dependent hawkes process for predicting retweet dynamics. In *Tenth International AAAI Conference on Web and Social Media*, March 2016.

[76] Harald Schmidbauer, Angi Rösch, and Fabian Stieler. The 2016 US presidential election and media on Instagram: Who was in the lead? *Computers in Human Behavior*, 81:148–160, April 2018.

[77] Philip N. Howard, Aiden Duffy, Deen Freelon, M. M. Hussain, Will Mari, and Marwa Maziad. Opening closed regimes: What was the role of social media during the Arab Spring? SSRN Scholarly Paper ID 2595096, Social Science Research Network, Rochester, NY, 2011.

[78] Zeynep Tufekci and Christopher Wilson. Social media and the decision to participate in political protest: Observations from Tahrir Square. *Journal of Communication*, 62(2):363–379, April 2012.

[79] Damian J. Ruck, Natalie M. Rice, Joshua Borycz, and R. Alexander Bentley. Internet Research Agency Twitter activity predicted 2016 U.S. election polls. *First Monday*, 24(7), June 2019.

[80] R. Kelly Garrett. Social media's contribution to political misperceptions in U.S. Presidential elections. *PLOS ONE*, 14(3), March 2019.

[81] Andrew Guess, Jonathan Nagler, and Joshua Tucker. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, 5(1), January 2019.

[82] Darko Cherepnalkoski, Andreas Karpf, Igor Mozetič, and Miha Grčar. Cohesion and coalition formation in the European Parliament: Roll-call votes and Twitter activities. *PLOS ONE*, 11(11), November 2016.

[83] Pablo Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1):76–91, 2015.

[84] Sarah J. Jackson and Brooke Foucault Welles. Hijacking #MYNYPD: Social Media Dissent and Networked Counterpublics. *Journal of Communication*, 65(6):932–952, December 2015.

[85] Sarah J. Jackson and Brooke Foucault Welles. #Ferguson is everywhere: Initiators in emerging counterpublic networks. *Information, Communication & Society*, 19(3):397–418, March 2016.

[86] Yannis Theocharis, Will Lowe, Jan W. van Deth, and Gema García-Albacete. Using Twitter to mobilize protest action: Online mobilization patterns and action repertoires in the Occupy Wall Street, Indignados, and Aganaktismenoi movements. *Information, Communication & Society*, 18(2):202–220, February 2015.

[87] Alexandre Bovet and Hernán A. Makse. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, 10(1):1–14, January 2019.

[88] Jayeon Lee and Youngshin Lim. Gendered campaign tweets: The cases of Hillary Clinton and Donald Trump. *Public Relations Review*, 42(5):849–855, December 2016.

[89] Kareem Darwish, Walid Magdy, and Tahar Zanouda. Trump vs. Hillary: What went viral during the 2016 US presidential election. In Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri, editors, *Social Informatics*, Lecture Notes in Computer Science, pages 143–161, Cham, 2017. Springer International Publishing.

[90] John Quelch and Thales Teixeira. The Twitter Election. *HBS Working Knowledge*, September 2016.

[91] Isobelle Clarke and Jack Grieve. Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. *PLOS ONE*, 14(9), September 2019.

[92] Sara Ahmadian, Sara Azarshahi, and Delroy L. Paulhus. Explaining Donald Trump via communication style: Grandiosity, informality, and dynamism. *Personality and Individual Differences*, 107:49–53, March 2017.

[93] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, March 2011.

[94] Bloomberg. Bloomberg launches a Twitter feed optimized for trading | Bloomberg L.P. July 2018.

[95] Ahmed Abdeen Hamed, Xindong Wu, Robert Erickson, and Tamer Fandy. Twitter K-H networks in action: Advancing biomedical literature for drug search. *Journal of Biomedical Informatics*, 56:157–168, August 2015.

[96] Nicole Wetsman. How Twitter is changing medical research. *Nature Medicine*, December 2019.

[97] Emily M. Cody, Andrew J. Reagan, Peter Sheridan Dodds, and Christopher M. Danforth. Public opinion polling with Twitter. *arXiv:1608.02024 [physics]*, August 2016. arXiv: 1608.02024.

[98] Alexandre Bovet, Flaviano Morone, and Hernán A. Makse. Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump. *Scientific Reports*, 8(1):1–16, June 2018.

[99] Ussama Yaqub, Soon Ae Chun, Vijayalakshmi Atluri, and Jaideep Vaidya. Analysis of political discourse on Twitter in the context of the 2016 US presidential elections. *Government Information Quarterly*, 34(4):613–626, December 2017.

[100] Joseph DiGrazia, Karissa McKelvey, Johan Bollen, and Fabio Rojas. More tweets, more votes: Social media as a quantitative indicator of political behavior. *PLOS ONE*, 8(11), November 2013.

[101] Josh Pasek, Colleen A. McClain, Frank Newport, and Stephanie Marken. Who's tweeting about the President? What big survey data can tell us about digital traces? *Social Science Computer Review*, page 0894439318822007, January 2019.

[102] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth International AAAI Conference on Weblogs and Social Media*, May 2010.

[103] Yu Wang, Jiebo Luo, Richard Niemi, Yuncheng Li, and Tianran Hu. Catching fire via "Likes": Inferring topic preferences of Trump followers on Twitter. In *Tenth International AAAI Conference on Web and Social Media*, March 2016.

[104] Cindy Hui, Yulia Tyshchuk, William A. Wallace, Malik Magdon-Ismail, and Mark Goldberg. Information cascades in social media in response to a crisis: A preliminary model and a case study. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, pages 653–656, Lyon, France, April 2012. Association for Computing Machinery.

[105] Raquel A. Baños, Javier Borge-Holthoefer, and Yamir Moreno. The role of hidden influentials in the diffusion of online information cascades. *EPJ Data Science*, 2(1):6, July 2013.

[106] Devipsita Bhattacharya and Sudha Ram. Sharing news articles using 140 characters: A diffusion analysis on Twitter. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 966–971, August 2012.

[107] Ashish Goel, Kamesh Munagala, Aneesh Sharma, and Hongyang Zhang. A note on modeling retweet cascades on Twitter. In David F. Gleich, Júlia Komjáthy, and Nelly Litvak, editors, *Algorithms and Models for the Web Graph*, Lecture Notes in Computer Science, pages 119–131, Cham, 2015. Springer International Publishing.

[108] Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, October 2008.

[109] Cristian Candia, C. Jara-Figueroa, Carlos Rodriguez-Sickert, Albert-László Barabási, and César A. Hidalgo. The universal decay of collective memory and attention. *Nature Human Behaviour*, 3(1):82–91, January 2019.

[110] Rafael Prieto Curiel, Carmen Cabrera Arnau, Mara Torres Pinedo, Humberto González Ramírez, and Steven R. Bishop. Temporal and spatial analysis of the media spotlight. *Computers, Environment and Urban Systems*, 75:254–263, May 2019.

[111] Giambattista Amati, Simone Angelini, Francesca Capri, Giorgio Gambosi, Gianluca Rossi, and Paola Vocca. Modelling the temporal evolution of the retweet graph. *IADIS International Journal on Computer Science & Information Systems*, 11(2), 2016.

[112] Marijn ten Thij, Tanneke Ouboter, Daniël Worm, Nelly Litvak, Hans van den Berg, and Sandjai Bhulai. Modelling of trends in Twitter using retweet graph dynamics. In Anthony Bonato, Fan Chung Graham, and Paweł Prałat, editors, *Algorithms and Models for the Web Graph*, Lecture Notes in Computer Science, pages 132–147, Cham, 2014. Springer International Publishing.

[113] Soumajit Pramanik, Qinna Wang, Maximilien Danisch, Jean-Loup Guillaume, and Bivas Mitra. Modeling cascade formation in Twitter amidst mentions and retweets. *Social Network Analysis and Mining*, 7(1):41, August 2017.

[114] David Kempe, Jon Kleinberg, and Eva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 137–146, Washington, D.C., August 2003. Association for Computing Machinery.

[115] Hongshan Jin, Masashi Toyoda, and Naoki Yoshinaga. Can cross-lingual information cascades be predicted on Twitter? In Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri, editors, *Social Informatics*, Lecture Notes in Computer Science, pages 457–472, Cham, 2017. Springer International Publishing.

[116] Kyumin Lee, Jalal Mahmud, Jilin Chen, Michelle Zhou, and Jeffrey Nichols. Who Will Retweet This? Detecting Strangers from Twitter to Retweet In-

formation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3):31:1–31:25, April 2015.

[117] Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Who says what to whom on Twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 705–714, Hyderabad, India, March 2011. Association for Computing Machinery.

[118] Pablo Barberá and Gonzalo Rivero. Understanding the political representativeness of Twitter users. *Social Science Computer Review*, December 2014.

[119] Jonathan Mellon and Christopher Prosser. Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, 4(3):2053168017720008, July 2017.

[120] Dhiraj Murthy, Alexander Gross, and Alexander Pensavalle. Urban social media demographics: An exploration of Twitter use in major American cities. *Journal of Computer-Mediated Communication*, 21(1):33–49, January 2016.

[121] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on Twitter during the 2016 U.S. presidential election. *Science*, 363(6425):374–378, January 2019.

[122] Michael D. Shear, Maggie Haberman, Nicholas Confessore, Karen Yourish, Larry Buchanan, and Keith Collins. How Trump Reshaped the Presidency in Over 11,000 Tweets. *The New York Times*, November 2019.

[123] Peter Sheridan Dodds, Joshua R. Minot, Michael V. Arnold, Thayer Alshaabi, Jane Lydia Adams, David Rushing Dewhurst, Andrew J. Reagan, and Christopher M. Danforth. Fame and Ultrafame: Measuring and comparing daily levels of 'being talked about' for United States' presidents, their rivals, God, countries, and K-pop. *arXiv:1910.00149 [physics]*, September 2019. arXiv: 1910.00149.

[124] Sarah J. Jackson, Moya Bailey, Brooke Foucault Welles, and Genie Lauren. *#HashtagActivism: Networks of Race and Gender Justice*. The MIT Press, Cambridge, March 2020.

[125] Twitter. Decahose stream. https://developer.twitter.com/en/docs/tweets/sample-realtime/overview/decahose, 2019.

[126] Twitter. Twitter IDs (snowflake). https://developer.twitter.com/en/docs/basics/twitter-ids, 2019.

[127] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12), 2011.

[128] Mohsen Sayyadiharikandeh, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2725–2732, 2020.

[129] Faiza Khan Khattak, Serena Jeblee, Chloé Pou-Prom, Mohamed Abdalla, Christopher Meaney, and Frank Rudzicz. A survey of word embeddings for clinical text. *Journal of Biomedical Informatics: X*, 4:100057, 2019.

[130] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning – a brief history, state-of-the-art and challenges, 2020.

[131] Katarina Hamberg. Gender bias in medicine. *Women's Health*, 4(3):237–243, 2008.

[132] Oras A Alabas, Chris P Gale, Marlous Hall, Mark J Rutherford, Karolina Szummer, Sofia Sederholm Lawesson, Joakim Alfredsson, Bertil Lindahl, and Tomas Jernberg. Sex differences in treatments, relative survival, and excess mortality following acute myocardial infarction: national cohort study using the swedeheart registry. *Journal of the American Heart Association*, 6(12):e007123, 2017.

[133] Brad N Greenwood, Seth Carnahan, and Laura Huang. Patient–physician gender concordance and increased mortality among female heart attack patients. *Proceedings of the National Academy of Sciences*, 115(34):8569–8574, 2018.

[134] Paul M Galdas, Francine Cheater, and Paul Marshall. Men and health help-seeking behaviour: literature review. *Journal of advanced nursing*, 49(6):616–623, 2005.

[135] Revital Gross, Rob McNeill, Peter Davis, Roy Lay-Yee, Santosh Jatrana, and Peter Crampton. The association of gender concordance and primary care physicians' perceptions of their patients. *Women & Health*, 48(2):123–144, 2008.

[136] Jyoti Malhotra, David Rotter, Jennifer Tsui, Adana AM Llanos, Bijal A Balasubramanian, and Kitaw Demissie. Impact of patient–provider race, ethnicity, and gender concordance on cancer screening: Findings from medical expenditure panel survey. *Cancer Epidemiology and Prevention Biomarkers*, 26(12):1804–1811, 2017.

[137] Brad N Greenwood, Rachel R Hardeman, Laura Huang, and Aaron Sojourner. Physician–patient racial concordance and disparities in birthing mortality for newborns. *Proceedings of the National Academy of Sciences*, 117(35):21194–21200, 2020.

[138] Marcella Alsan, Owen Garrick, and Grant C Graziani. Does diversity matter for health? experimental evidence from oakland. Technical report, National Bureau of Economic Research, 2018.

[139] Marcella Alsan and Marianne Wanamaker. Tuskegee and the health of black men. *The quarterly journal of economics*, 133(1):407–455, 2018.

[140] Chi-Wei Lin, Meei-Ju Lin, Chin-Chen Wen, and Shao-Yin Chu. A word-count approach to analyze linguistic patterns in the reflective writings of medical students. *Medical education online*, 21(1):29522, 2016.

[141] Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, 2019.

[142] Víctor M Prieto, Sergio Matos, Manuel Alvarez, Fidel Cacheda, and José Luís Oliveira. Twitter: a good place to detect health conditions. *PloS one*, 9(1):e86191, 2014.

[143] Manuel Rodríguez-Martínez and Cristian C Garzón-Alfonso. Twitter health surveillance (ths) system. In *Proceedings:... IEEE International Conference on Big Data. IEEE International Conference on Big Data*, volume 2018, page 1647. NIH Public Access, 2018.

[144] Marcel Salathé. Digital epidemiology: what is it, and where is it going? *Life sciences, society and policy*, 14(1):1–5, 2018.

[145] Michelle M Mello and C Jason Wang. Ethics and governance for digital disease surveillance. *Science*, 368(6494):951–954, 2020.

[146] Justin Sybrandt, Michael Shtutman, and Ilya Safro. Moliere: Automatic biomedical hypothesis generation system. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1633–1642, 2017.

[147] Christopher C Yang, Haodong Yang, Ling Jiang, and Mi Zhang. Social media mining for drug safety signal detection. In *Proceedings of the 2012 international workshop on Smart health and wellbeing*, pages 33–40, 2012.

[148] Bethany Percha, Yuhao Zhang, Selen Bozkurt, Daniel Rubin, Russ B Altman, and Curtis P Langlotz. Expanding a radiology lexicon using contextual patterns in radiology reports. *Journal of the American Medical Informatics Association*, 25(6):679–685, 2018.

[149] Yadan Fan, Serguei Pakhomov, Reed McEwan, Wendi Zhao, Elizabeth Lindemann, and Rui Zhang. Using word embeddings to expand terminology of dietary supplements on clinical notes. *JAMIA open*, 2(2):246–253, 2019.

[150] Frederic Morin and Yoshua Bengio. Hierarchical probabilistic neural network language model. In *Aistats*, volume 5, pages 246–252. Citeseer, 2005.

[151] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[152] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[153] Iz Beltagy, Arman Cohan, and Kyle Lo. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[154] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[155] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[156] Kexin Huang, Abhishek Singh, Sitong Chen, Edward T Moseley, Chih-ying Deng, Naomi George, and Charlotta Lindvall. Clinical xlnet: Modeling sequential clinical notes and predicting prolonged mechanical ventilation. *arXiv preprint arXiv:1912.11975*, 2019.

[157] Marion Bartl, Malvina Nissim, and Albert Gatt. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. *arXiv preprint arXiv:2010.14534*, 2020.

[158] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*, 2019.

[159] Christine Basta, Marta R. Costa-jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy, August 2019. Association for Computational Linguistics.

[160] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in Neural Information Processing Systems*, 33, 2020.

[161] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

[162] Erenay Dayanik and Sebastian Padó. Masking actor information leads to fairer political claims detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4385–4391, 2020.

[163] Bo Liu. Anonymized bert: an augmentation approach to the gendered pronoun resolution challenge. *arXiv preprint arXiv:1905.01780*, 2019.

[164] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, 2019.

[165] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

[166] Sosuke Kobayashi. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*, 2018.

[167] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.

[168] Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10(1):1–16, 2010.

[169] Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606, 2017.

[170] Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

[171] David Ifeoluwa Adelani, Ali Davody, Thomas Kleinbauer, and Dietrich Klakow. Privacy guarantees for de-identifying text transformations. *arXiv preprint arXiv:2008.03101*, 2020.

[172] Thang M Pham, Trung Bui, Long Mai, and Anh Nguyen. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? *arXiv preprint arXiv:2012.15180*, 2020.

[173] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, 2020.

[174] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

[175] Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 2004.

[176] Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining*, 14(1):1–22, 2021.

[177] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*, 2019.

[178] Rishi Bommasani, Kelly Davis, and Claire Cardie. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, 2020.

[179] Vishesh Kumar, Amber Stubbs, Stanley Shaw, and Özlem Uzuner. Creation of a new longitudinal corpus of clinical narratives. *Journal of biomedical informatics*, 58:S6–S10, 2015.

[180] Haggi Mazeh, Rebecca S Sippel, and Herbert Chen. The role of gender in primary hyperparathyroidism: same disease, different presentation. *Annals of surgical oncology*, 19(9):2958–2962, 2012.

[181] Yang Trista Cao and Hal Daumé III. Toward gender-inclusive coreference resolution. *arXiv preprint arXiv:1910.13913*, 2019.

[182] Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*, 2019.

[183] Cayla R Teal, Anne C Gill, Alexander R Green, and Sonia Crandall. Helping medical learners recognise and manage unconscious bias toward certain patient groups. *Medical education*, 46(1):80–88, 2012.

[184] Javier Sánchez-Monedero, Lina Dencik, and Lilian Edwards. What does it mean to solve the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems. *CoRR*, abs/1910.06144, 2019.

[185] Ifeoma Ajunwa, Sorelle Friedler, Carlos E Scheidegger, and Suresh Venkatasubramanian. Hiring by algorithm: predicting and preventing disparate impact. *Available at SSRN*, 2016.

[186] Rebecca Heilweil. Artificial intelligence will help determine if you get your next job — vox.com. https://www.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen. [Accessed 04-Apr-2022].

[187] Ketki V Deshpande, Shimei Pan, and James R Foulds. Mitigating demographic bias in ai-based resume filtering. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 268–275, 2020.

[188] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[189] Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17, 2019.

[190] Miranda Bogen and Aaron Rieke. Help wanted: An examination of hiring algorithms, equity, and bias. *Upturn*, 2018.

[191] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 469–481, 2020.

[192] Liangyue Li, How Jing, Hanghang Tong, Jaewon Yang, Qi He, and Bee-Chung Chen. Nemo: Next career move prediction with contextual embedding. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 505–513, 2017.

[193] Ahmad Alabdulkareem, Morgan R Frank, Lijun Sun, Bedoor AlShebli, César Hidalgo, and Iyad Rahwan. Unpacking the polarization of workplace skills. *Science advances*, 4(7):eaao6030, 2018.

[194] Denghui Zhang, Junming Liu, Hengshu Zhu, Yanchi Liu, Lichen Wang, Pengyang Wang, and Hui Xiong. Job2vec: Job title benchmarking with collective multi-view representation learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2763–2771, 2019.

[195] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018.

[196] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[197] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China, November 2019. Association for Computational Linguistics.

[198] Hannah Rose Kirk, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, Yuki Asano, et al. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in Neural Information Processing Systems*, 34, 2021.

[199] Claire Cain Miller. Job Listings That Are Too 'Feminine' for Men. *The New York Times*, January 2017.

[200] Nicole Lewis. Will ai remove hiring bias. *SHRM*, 2018.

[201] Martijn Wieling, John Nerbonne, and R Harald Baayen. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PloS one*, 6(9):e23613, 2011.

[202] Jacob Levy Abitbol, Márton Karsai, Jean-Philippe Magué, Jean-Pierre Chevrot, and Eric Fleury. Socioeconomic dependencies of linguistic patterns in twitter: A multivariate analysis. In *Proceedings of the 2018 World Wide Web Conference*, pages 1125–1134, 2018.

[203] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205, 2006.

[204] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[205] Anders Johannsen, Dirk Hovy, and Anders Søgaard. Cross-lingual syntactic variation over age and gender. In *Proceedings of the nineteenth conference on computational natural language learning*, pages 103–112, 2015.

[206] Carol Isaac, Jocelyn Chertoff, Barbara Lee, and Molly Carnes. Do students' and authors' genders affect evaluations? a linguistic analysis of medical student performance evaluations. *Academic Medicine*, 86(1):59, 2011.

[207] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. In *Proceedings of NAACL-HLT*, pages 629–634, 2019.

[208] Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, April 2022.

[209] Jayadev Bhaskaran and Isha Bhallamudi. Good secretaries, bad truck drivers? occupational gender stereotypes in sentiment analysis. *GeBNLP 2019*, page 62, 2019.

[210] Dirk Hovy. Demographic factors improve classification performance. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 752–762, 2015.

[211] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[212] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[213] Joshua R Minot, Nicholas Cheney, Marc Maier, Danne C Elbers, Christopher M Danforth, and Peter Sheridan Dodds. Interpretable bias mitigation for textual data: Reducing genderization in patient notes while maintaining classification performance. *ACM Transactions on Computing for Healthcare*, 2022.

[214] Amanda Barroso and Anna Brown. Gender pay gap in us held steady in 2020. *Pew Research Center*, 2021.

[215] Francine D Blau and Lawrence M Kahn. The gender wage gap: Extent, trends, and explanations. *Journal of economic literature*, 55(3):789–865, 2017.

[216] Asaf Levanon, Paula England, and Paul Allison. Occupational feminization and pay: Assessing causal dynamics using 1950–2000 us census data. *Social forces*, 88(2):865–891, 2009.

[217] Gaëlle Ferrant, Luca Maria Pesando, and Keiko Nowacka. Unpaid care work: The missing link in the analysis of gender gaps in labour outcomes. *Boulogne Billancourt: OECD Development Center*, 2014.

[218] Sherry E Sullivan and Lisa A Mainiero. The changing nature of gender roles, alpha/beta careers and work-life issues: Theory-driven implications for human resource management. *Career Development International*, 2007.

[219] Megan Dunn. Who chooses part-time work and why. *Monthly Lab. Rev.*, 141:1, 2018.

[220] Allison C Morgan, Samuel F Way, Michael JD Hoefer, Daniel B Larremore, Mirta Galesic, and Aaron Clauset. The unequal impact of parenthood in academia. *Science Advances*, 7(9):eabd1996, 2021.

[221] Richard J Petts, Daniel L Carlson, and Joanna R Pepin. A gendered pandemic: Childcare, homeschooling, and parents' employment during covid-19. *Gender, Work & Organization*, 28:515–534, 2021.

[222] Barbara Petrongolo and Maddalena Ronchi. Gender gaps and the structure of local labor markets. *Labour Economics*, 64:101819, 2020.

[223] Sapna Cheryan, Sianna A Ziegler, Amanda K Montoya, and Lily Jiang. Why are some stem fields more gender balanced than others? *Psychological bulletin*, 143(1):1, 2017.

[224] Clotilde Napp and Thomas Breda. The stereotype that girls lack talent: A worldwide investigation. *Science advances*, 8(10):eabm3689, 2022.

[225] Shelley J Correll, Stephen Benard, and In Paik. Getting a job: Is there a motherhood penalty? *American journal of sociology*, 112(5):1297–1338, 2007.

[226] Ewa Cukrowska-Torzewska and Anna Lovasz. The role of parenthood in shaping the gender wage gap–a comparative analysis of 26 european countries. *Social Science Research*, 85:102355, 2020.

[227] Ali Ahmed, Mark Granberg, and Shantanu Khanna. Gender discrimination in hiring: An experimental reexamination of the swedish case. *PloS one*, 16(1):e0245513, 2021.

[228] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14, 2018.

[229] David Deming and Lisa B Kahn. Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics*, 36(S1):S337–S369, 2018.

[230] Sarah H Bana. job2vec: Using language models to understand wage premia. 2021.

[231] Michael S Cole, Hubert S Feild, and William F Giles. Interaction of recruiter and applicant gender in resume evaluation: a field study. *Sex Roles*, 51(9):597–608, 2004.

[232] Robert L Dipboye, Howard L Fromkin, and Kent Wiback. Relative importance of applicant sex, attractiveness, and scholastic standing in evaluation of job applicant resumes. *Journal of Applied Psychology*, 60(1):39, 1975.

[233] Randall L Popken. An empirical study of indirect speech acts in resumes. *English for Specific Purposes*, 12(3):239–249, 1993.

[234] Bureau of Labor Statistics, U.S. Department of Labor. Standard Occupational Classification (SOC) System — bls.gov. https://www.bls.gov/soc/2018/#classification. [Accessed 03-Apr-2022].

[235] National Center for O*NET Development. O*net online. www.onetonline.org/. [Accessed 04-May-2022].

[236] Nils Reimers and Iryna Gurevych. 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, 11 2019.

[237] Lucía Santamaría and Helena Mihaljević. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, 4:e156, 2018.

[238] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[239] Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. *arXiv preprint arXiv:2205.04610*, 2022.

[240] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. Bayesian modeling of intersectional fairness: The variance of bias. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 424–432. SIAM, 2020.