Abstract

Algorithmic Decision-Making with Stakeholder Participation

Vijay Keswani

2023

The development of trustworthy systems for applications of machine learning and artificial intelligence faces a variety of challenges. These challenges range from the investigation of methods to effectively detect algorithmic biases to methodological and practical hurdles encountered when incorporating notions of representation, equality, and domain expertise in automated decisions. Such questions make the task of building reliable automated decision-making frameworks quite complex; nevertheless, addressing them in a comprehensive manner is an important step toward building automated tools whose impact is equitable. This dissertation focuses on tackling such practical issues faced during the implementation of automated decision-making frameworks. It contributes to the growing literature on algorithmic fairness and human-computer interaction by suggesting methods to develop frameworks that account for algorithmic biases and that encourage stakeholder participation in a principled manner.

I start with the problem of representation bias audit, i.e., determining how well a given data collection represents the underlying population demographics. For data collection from real-world sources, individual-level demographics are often unavailable, noisy, or restricted for automated usage. Employing user-specified representative examples, this dissertation proposes a cost-effective algorithm to approximate the representation disparity of any unlabeled data collection using the given examples. By eliciting examples from the users, this method incorporates the users' notions of diversity and informs them of the extent to which the given data collection under or over-represents socially-salient groups. User-defined rep-

resentative examples are further used to improve the diversity of automatically-generated summaries for text and image data collections, ensuring that the generated summaries appropriately represent all relevant groups.

The latter part of the dissertation studies the paradigm of human-in-the-loop deferral learning. In this setting, the decision-making framework is trained to either make an accurate prediction or defer to a domain expert in cases where the algorithm has low confidence in its inference. Our work proposes methods for training a deferral framework when multiple domain experts are available to assist with decision-making. Using appropriate statistical fairness mechanisms, the framework ensures that the final decisions maintain performance parity across demographic groups.

By focusing on stakeholder participation, in the forms of user feedback incorporation or domain expert participation, this dissertation advances methods to build trustworthy decision-making systems which can be readily deployed in practice.

Algorithmic Decision-Making with Stakeholder Participation

A Dissertation

Presented to the Faculty of the Graduate School

of

Yale University

in Candidacy for the Degree of

Doctor of Philosophy

by

Vijay Keswani

Dissertation Director: L. Elisa Celis

May 2023

Contents

Acknowledgements

To Elisa Celis, who has been the best advisor I could have asked for. Your passion for the subject of data ethics and your engagement with this field beyond computer and data sciences have been deeply inspiring. You taught me the importance of clarity and curiosity in creating meaningful research and encouraged me to explore topics beyond my comfort zone. Even when this led to me pursuing unconventional projects, you still always supported me and worked with me to refine my ideas. I am truly grateful for the faith you have shown in me.

To Nisheeth K. Vishnoi, who has been a mentor and a guide to me throughout my Ph.D. Working with you helped me understand the importance of rigor and perseverance in research. You have always emphasized the significance of asking the right questions, developing a better theoretical understanding of my research, and communicating my work in a clear and concise manner. Your lessons have been paramount in my research.

To Matthew Lease and Krishnaram Kenthapadi, thank you for being my mentors and collaborators. I appreciate your patience, encouragement to continuously pursue difficult research problems, and advice whenever I needed it. It has been a privilege working with you both over the last few years.

Thank you to Jas Sekhon and Matthew Lease for being the readers of this dissertation and to Karen Kavanaugh, Jay Emerson, and Andrew Barron for all the administrative and academic help.

To Chinmayi Arun at the Information Society Project and Demar Lewis and Chloe Sariego at the Institute of Social and Policy Studies, thank you for having me as a fellow at your institutes. Being a part of these communities helped me develop a better appreciation for the complexity that comes along with questions of ethics and I am grateful to these institutes and the affiliated scholars for providing me with opportunities to grow beyond the field of data science.

To my fellow Ph.D. colleagues, Colleen, Anay, Curtis, Alex, Shinpei, Megan, and Sky, who kept me academically afloat, helped me troubleshoot my problems, proofread this thesis, gave me amazing feedback on my research and writing, and always patiently listened to my rants about research and beyond. I consider myself incredibly lucky to have colleagues and friends as supportive as you.

To Stacey, Sarah, Doug, Halley, Topaz, Garth, Prabaha, Vasudha, Kabish, Akshay, and Fiona, friends and housemates whose company kept me sane through difficult days. My time in New Haven has been lovely, enriching, and full of pleasant surprises thanks to all of you and I will cherish the moments we spent together. Last but not the least, I thank my family for their neverending love and humor. I am sure my decisions don't always make sense to you but I will forever be grateful for your endless support. I am who I am thanks to all of you.

Chapter 1

Introduction

A rational decision-making process incorporates a variety of values and preferences of the decision-maker. Using information from prior decisions, evaluation of counterfactuals, and ranking available actions by priority, decision-making involves a complex mechanism that we, as humans, execute habitually. Our decisions express our personal and social preferences and embody the values we deem important. Yet, our decision-making processes are not perfect and we all face moments where our decisions are incorrect. These failures can stem from inadequate prior information, lack of experience, or from other internal and external factors. Considering the impossibility of any one human possessing all the knowledge and experience in the world, we rely on each other to make correct decisions. We defer, we ask for help, and we learn from others to improve our decision-making. We develop automated tools like computers to assist our decision-making by using them for routine tasks like arithmetic computations or by employing them for complex tasks that involve advanced algorithmic systems like map navigation.

A crucial aspect of any decision-making process that involves two or more parties is trust. The exercise of trust building involves beliefs of shared values and interests among the decision-makers and facilitates the acceptance of one party's

decision by the other. With the involvement of automated tools in our decision-making process, the question of trust comes up time and again. *Do we trust automated tools to make decisions that embody our values? Do we trust automated tools to account for our preferences in an objective manner? Do we trust automated tools to make decisions in a way that would be the most beneficial to us?* We trust other humans to assist our decisions when they have demonstrated, through intent and action, that they share similar interests as us. *Can we place the same trust in automated tools that we didn't develop and in algorithms that we didn't design which are, nevertheless, parts of our daily lives?* This dissertation explores this question by investigating the decisions made by Artificial Intelligence (AI) and Machine Learning (ML) tools through the perspectives of users and stakeholders. I demonstrate how flawed algorithmic mechanisms can lead to harmful automated decisions and design methods to counter algorithmic harms, whenever possible, through a judicious process by which stakeholders are a part of the algorithmic decision-making process.

The availability of large datasets, massive computing power, and progress in machine learning methods has led to a surge in the use of automated decision-making frameworks in a variety of domains. Technological and monetary investments have facilitated significant improvements in the performance of algorithmic tools and a number of applications of these tools lie in fields that make decisions affecting humans and society in general. They are employed in numerous critical applications, including healthcare [152, 286, 302], advertising [245, 257], online search and recommendation feeds [40, 180], lending [231, 324], content moderation [80, 220, 313], recruitment [106], criminal risk assessment [1, 89, 128, 233], and policing [134, 149]. All these applications involve actively processing information related to people and making decisions that affect society at an individual and institutional level. The impact of such automated frameworks in shaping our current and future socio-technical landscape cannot be understated.

A technical taxonomy of applications of artificial intelligence involves considering different kinds of learning methodologies involved in the applications. In this context, two popular learning approaches that cover a large number of applications are unsupervised and supervised learning[1]. Unsupervised learning corresponds to processing large amounts of unlabelled data to extract useful structural and semantic information about the data [48, 82]. For instance, the task of clustering or ranking a large set of images to generate a small subset of representative images (e.g., ranking in search engines or recommendation feeds) is a prominent use case of unsupervised learning. Supervised learning, on the other hand, is used to develop labeling or prediction algorithms that can predict task-relevant outcomes for given data points [47, 140]. Supervised learning techniques are used to train decision-making frameworks on outcome-labeled datasets, with the goal of accurately predicting the outcomes for future data. For example, past human hiring decisions can be used to train an automated recruitment pipeline that then makes hiring decisions for future applicants. Both learning paradigms are widely employed in a variety of domains. Unsupervised learning tools, such as recommendation and search, fulfill important informational gaps between data and the underlying population structure and are now an integral part of our interaction with the digital world. Similarly, supervised learning algorithms, such as classification and regression, are trained to simulate past decisions and deployed to assist future decision-making.

Trust in these automated tools is usually established by testing them on real-world scenarios and quantifying their performance using statistical measures. While some errors are expected (as in the case of even human decision-making), we nev-

[1]There are other learning paradigms as well, including reinforcement learning and a spectrum of semi-supervised learning methods that combine techniques from supervised and unsupervised learning. For this thesis, the focus on supervised and unsupervised learning arises from the interest in specific applications where algorithmic harms are commonly encountered. See Section 2.2 for further discussion.

ertheless demand these automated tools to demonstrate that their decisions align with the users' preferences. For instance, search engines sometimes return results that do not provide us with the information we are looking for. Random errors might be excusable if they occur infrequently and the overall decision accuracy is sufficiently high; however, systemic errors that reflect problematic decision-making patterns reveal deeper issues with the use of automation. Investigating the pattern of decisions made by certain automated tools indeed paints a grim picture: real-world algorithmic decisions often encode problematic social biases and disparately favor some demographic groups over others. Furthermore, this disparity mirrors the divide in our society as algorithms employed in real-world practice exhibit and even propagate societal inequalities and negative stereotypes against groups that have been historically disadvantaged. In the case of unsupervised learning, real-world applications of summarization or retrieval algorithms have been shown to exhibit gender and racial biases, leading to a stereotypically-biased representation of underlying populations [166, 232]. Similarly, supervised algorithms deployed in practice often have disparate performance for different demographic groups, such as in the settings of criminal recidivism [1, 64, 94, 182], predictive policing [103, 149, 259, 272], recruitment [78, 264], and healthcare [92, 235]. Clearly, the presence of these biases undermines the trust we can place in the decisions of automated tools. Correspondingly, it is important to study methods that can (a) evaluate the biases in automated tools, and (b) if possible, modify these tools so that they do not inherit and propagate social biases of the data or the developers. Chapter 2 presents an overview of the research on social biases, popular techniques for automated decision-making, and prior studies demonstrating social biases in automated decisions.

Addressing social biases in algorithmic tools requires overcoming many different kinds of challenges. From a practical viewpoint, the definition of what one

considers to be *unbiased* or *fair* is highly context-dependent and relies crucially on the stakeholders involved in the design of the framework. From a technical viewpoint, ensuring that the output of an algorithm is *fair* with respect to socially salient attributes, such as gender, race, skintone, or dialect, often requires incorporating additional constraints or posthoc adjustments into the learning process, making the task of learning the final framework quite complex [173]. These practical and technical challenges manifest themselves in different ways in different applications, making the process of bias mitigation a highly involved exercise that requires the participation of both users and designers of the framework to converge to an accurate and equitable decision-making framework. This dissertation discusses both challenges using the methodological frameworks of popular applications of AI, such as Google Image Search, Twitter recommendation feeds, and human-AI teams for content moderation. In all of these applications, the presented research studies the impact of social biases, suggests methods to audit them efficiently, and, in most cases, proposes solutions that can function as unbiased alternatives in these applications. The proposed solutions take into account the hurdles one can encounter while implementing these frameworks in real-world settings and aims to provide feasible solutions to address biases despite such hurdles.

The first step towards addressing biases in any algorithmic application is to develop methods to efficiently detect or audit them. The statistical question of bias audit essentially boils down to employing hypothesis testing frameworks to determine if there are disparities in the representation of different groups in any given data collection. However, this simple process of bias audit becomes difficult to implement when the group memberships or socially salient attributes (e.g., gender or skintone) of individual samples are unknown[2]. For example, suppose we wish to

[2]I will use terms socially salient attributes and protected attributes interchangeably throughout the dissertation. While protected attributes usually correspond to group identities that are protected by anti-discrimination laws, I will use this term to also denote attributes that we wish to protect against algorithmic harm. See Section 2.4 for further discussion on this point.

check the disparity in gender representation of Google Image Search results for any given occupation. Executing this task automatically is difficult since the presented gender of the people in the images is quite often unavailable. Auditing these results, in this case, would then involve manual labeling or crowdsourced labeling of the perceived gender, which can be expensive and time-consuming. Chapter 3 presents an alternative - i.e., an efficient algorithm for auditing representational biases in the absence of socially salient feature information. The proposed algorithm uses a small set of labeled representative examples (which can be user-specified) to measure representation disparity in any given unlabeled dataset, under certain domain assumptions. To measure representation disparity with respect to any socially salient attribute (i.e., the difference in the fraction of elements with one attribute value vs. another), this algorithm calculates the average similarity between the elements in the unlabeled dataset and the elements in the labeled set of representative examples. Using these similarity scores, we can approximate the representation disparity by taking the difference between group-wise similarity scores. Theoretical analysis using standard concentration inequalities demonstrates that the proposed algorithm produces a good approximation of the actual representation disparity of the dataset even when the number of labeled examples is logarithmic in the size of the unlabeled dataset. To further reduce the approximation error, we also propose an algorithm that can construct an *appropriate* set of labeled examples for auditing purposes. Empirical evaluations on multiple image and text datasets demonstrate that the proposed audit algorithm effectively approximates the representation disparity in any random or topic-specific data collection.

The primary contribution of the above bias audit algorithm is the use of representative examples. These user-defined representative samples incorporate the user's notion of diversity and side-step the issue of unavailable group attributes. We extend the use of such representative samples to debias automatically-generated

summaries. Chapters 4 and 5 cover the field of fair summarization and present post-processing algorithms for generating diverse summaries using a small set of representative examples. Both chapters first highlight the presence of social biases in the outputs of popular image and text summarization algorithms and then use suggest methods to improve group representations in automatically-generated summaries using user-defined representative examples. Chapter 4 focuses on image summarization, where we first evaluate the diversity in Google Image Search results. To do so, we collect top image search results using 96 occupations as search queries (extending the methodology of Kay et al. [166]). We observe that the search results consistently favor and over-represent gender-stereotypical and skintone-stereotypical images. Given this issue of misrepresentation, we next propose efficient methods to incorporate visible diversity in summary results using user-defined representative examples. Once again, note that these data collections can be at scales where collecting socially salient attributes or group labels is infeasible (e.g., search engine results for any possible query) and the use of representative examples can side-step the issue of unavailable attributes. We propose two post-processing algorithms, inspired by the well-known *Maximal Marginal Relevance* (MMR) algorithm [46], to debias image summaries in a post-processing manner. Our algorithms take a black-box image summarization algorithm and the unlabeled dataset to be summarized as input and overlay it with a post-processing step that diversifies the results of the black-box algorithm using the given representative examples. We demonstrate the efficacy of these algorithms over multiple image datasets, including the Google Image Search dataset we collected. For these datasets, we observe an improvement in demographic representation in generated summaries while ensuring that the summaries are visibly diverse in a similar manner as the user-defined representative examples.

Chapter 5 extends the use of our post-processing algorithm for the domain of

extractive text summarization, i.e., the task of generating a short summary for a large number of sentences. Again, we first demonstrate the lack of diversity in the summaries generated by popular extractive text summarization algorithms. In particular, our analysis considers diversity with respect to various dialects (e.g., Standard English and African-American English dialects) in datasets containing Twitter posts. We evaluate the dialect diversity in the summaries generated by frequency-based summarization algorithms (e.g., TF-IDF [203] and Hybrid TF-IDF [150]), graph-based algorithms (LexRank [104] and TextRank [209]), non-redundancy based algorithms (MMR [122] and Centroid-Word2Vec [262]), and pre-trained supervised approaches (SummaRuNNer [224]). We observe that, for random and topic-specific collections from these datasets, most algorithms return summaries that under-represent certain dialects. To address this dialect bias, we employ the post-processing algorithm from Chapter 4. As mentioned earlier, this approach requires a small set of representative labeled examples, which in this case is a small dialect-diverse set of Twitter posts given as part of the input. Using a small set of sentences written in different dialects as the set of representative examples, the post-processing algorithm efficiently increases the dialect diversity of any set of given Twitter posts, demonstrating the applicability of this approach for debiasing social media recommendation feeds.

Chapter 6 considers the supervised learning problem of training a decision-making framework given human assistance. In applications like risk assessment [127] and maltreatment hotline screening [65], multiple human experts are available to assist an automated decision-making framework, so as to share the load and to cover different kinds of input samples [129]. This chapter studies the setting where an automated decision-making framework can either make a prediction for a given input or defer the decision to a human expert when it has low confidence in its prediction. Since different human experts can have different domains of ex-

pertise and various social prejudices, choosing the appropriate unbiased expert when deferring the decision is crucial to ensure the high accuracy of final predictions. Hence, in this setting, there is an additional challenge of determining which decision-maker (among the available humans and the machine) should make the final decision. Chapter 6 presents a training framework that simultaneously learns an automated classifier and a deferral model, such that the classifier is the primary decision-maker but it defers the decision to an appropriate human for input subdomains where it lacks sufficient information. Theoretically, we show that this deferral framework can be trained efficiently using gradient descent-based methods and provide mechanisms to incorporate popular statistical fairness metrics with the deferral training. The efficacy of the framework is also demonstrated via synthetic experiments and real-world experiments, the latter conducted over a dataset we curate by asking a large number of crowd-annotators to label the toxicity of a collection of social media posts.

The methodologies presented in this dissertation focus on stakeholder participation. Chapters 3, 4, and 5 present algorithms that address biases using a user-specified representative set of examples. By utilizing these examples, we ensure that the final output of the framework aligns with the user's idea of diversity and create a participatory process to address representation biases. Similarly, Chapter 6 proposes methods to create decision-making frameworks that employ the available human experts in a manner that improves the overall predictive accuracy. Such a framework is most effective when the human experts are as diverse as the targeted user population. The inclusion of human feedback helps incorporate shared values, preferences, and expertise of the stakeholders. In this manner, the research in this dissertation aims to address crucial faults in the final decisions of automated decision-making frameworks using stakeholder participation, allowing us to steadily build trust in the decisions of these frameworks.

Chapter 2

Background

There has been significant interest in the field of fair machine learning and AI ethics in the last decade. Early investigations by journalists and academic scholars empirically demonstrated the presence of gender and racial biases in the outcomes of algorithmic frameworks [13, 105, 232, 211]. Seminal works by computer and data sciences researchers correspondingly studied methods to mathematically model these automated biases [22, 98, 136, 312]. Following the footsteps of these works and inspired by decades of research on decision-making biases in fields like sociology, law, philosophy, and economics, data science and computer science researchers have started critically assessing the biases present in different algorithmic applications. In this chapter, I present an overview of the research on social biases in automated decision-making and situate the work presented in this thesis within the larger fields of algorithmic fairness and human-computer interaction. Literature that is directly related to the research presented in this dissertation is relegated to the individual chapters. The discussion below starts with a brief introduction to the research on stereotypes and biases in human decision-making and then covers the relevant paradigms of automated decision-making and algorithmic fairness methods for machine learning and artificial intelligence applications.

2.1 Study of Stereotypes, Biases, and Their Impact

The study of cultivation and the impact of stereotypes has drawn serious interest in the age of digital media [237, 247], primarily due to the increased ease of information access and the possibility of stereotype propagation via sources like images on social media or search results. To define briefly, stereotyping is the process of inferring common characteristics of individuals in a group. When used accurately, stereotypes associated with a group are helpful in deducing information about individuals from the group in the absence of additional information [33, 207] and also function as tools to characterize group action [41, 139, 287]. However, inaccurate or exaggerated stereotypes can be quite harmful and can inadvertently cause biases against the individuals from the stereotyped group [116]. Prior studies have shown that the association of a negative stereotype with a group for a given task can affect the performance of the stereotyped individuals on the task [281, 306]; using the performance on such a task for any kind of future decision-making will lead to the propagation of such stereotypes and bias the results against one group. Furthermore, inaccurate stereotypes also lead to an incorrect perception of reality, especially with respect to sub-population demographics [117, 166, 275]. For example, stereotypical images of Black women as matriarchs or mammies, that are further disseminated via digital media, can lead to the normalization of such stereotypes [68, 138]. Given the existence of such negative social stereotypes and the possibility of their propagation via digital sources, it is important to explore methods to prevent their exacerbation through the use of automation.

The role of biases has seen similar investigation across social science disciplines. Decision-making biases often arise due to the decision-maker's prejudices against certain groups or due to a lack of information about individuals from certain groups (leading to a reliance on stereotypes) [24, 107]. These biases manifest themselves in the form of reduced access to resources or diminished performance

of decision-making systems for individuals from disadvantaged groups. Continuous audit of various human and institutional decision-making settings has revealed the presence of biases with respect to race, gender, and other demographic and socially-salient attributes in many common settings. This includes biases in socially-critical applications like mortgage approval [8], criminal justice system and policing [217], healthcare [92], recruitment [142], and social welfare access [277].

Frequent and extensive audits of these decision-making settings are crucial to ensure the accountability of the associated institutions. In particular, third-party audits of biases have been shown to be impactful in the past, often resulting in significant oversight and modification of harmful decision-making processes [4, 39]. It is important to subject automated decision-making to a similar level of continuous scrutiny and methods to efficiently audit or mitigate social biases can be useful in developing accountable and transparent technologies.

2.2 Automated Decision-Making

Automated decision-making can take a variety of forms and can be studied in the context of any application that involves machine support. For the purposes of this dissertation, I focus on automated frameworks that are designed to make decisions by processing large amounts of prior and current data and decisions.

As mentioned before, unsupervised learning algorithms learn mathematical (and potentially interpretable) patterns within a large data collection [48, 82]. Given a large number of samples from a particular domain, unsupervised learning algorithms aim to deduce the underlying representation of the samples which can then be used for future decision-making. Clustering, summarization, and outlier detection are all various instances of the unsupervised learning approach that allow for

a structured analysis of a large amount of data.

Supervised learning aims to learn the mathematical relationship between task-related features and the associated outcomes (usually characterized by *class labels*) through data [47, 140]. Given task-related features for samples observed in the past and the decisions made or true outcomes for these samples, supervised learning algorithms are used to infer a mathematical function that maps the features to the decisions/outcomes; this function can then be used to make decisions for future samples. The feature-decision pairs used to learn the function are called the *training data* for the learning algorithm. For example, in healthcare, this training data could correspond to health and demographic data of patients and whether they were afflicted with a particular disease. The supervised learning algorithm trained on this data can then be potentially used to predict the likelihood of any future patient suffering from the same disease using their health and demographic information.

The primary difference between supervised and unsupervised learning is that in unsupervised learning there are no "decisions" or labels associated with the available data. For example, clustering simply involves finding subsets within a given dataset such that elements within a subset are more similar to each other than to the elements outside the subset [178]. The learned cluster identities can then be used for downstream labeling or decision-making, but these identities wouldn't be known beforehand.

Finally, semi-supervised learning combines the paradigms of supervised and unsupervised learning and is applied in situations where a small amount of labeled data is available along with a large amount of unlabelled data. In this case, combining the function learned using the labeled data with representations learned using unlabelled data is important to build an overall robust decision-making system. Chapters 4 and 5 demonstrate the use of unsupervised and semi-supervised

learning paradigms for the task of summarization.

Note that automated decision-making is traditionally associated with just supervised or semi-supervised learning. This is because the notion of decision-making is clear in the applications of these paradigms – given data about past decisions, learn to simulate these decisions in the future. For unsupervised learning, prior decisions are not available. Nevertheless, the representations learned using unsupervised learning algorithms are still used for decision-making. Clustering algorithms are often used to identify the appropriate cluster for future samples so that cluster-specific processing techniques can be employed appropriately. Summarization algorithms are used to decide which samples best represent a given large collection. Recommendation systems similarly decide the content that is most likely to be relevant to a given user. Considering that the applications of unsupervised learning involve making automated decisions, I will use the term automated decision-making for unsupervised learning applications as well.

While the goal of supervised learning is to simulate (and potentially replace) human decision-making, in practice, automated decision-making tools are often deployed side-by-side with expert humans [84, 133]. For example, machine learning models in healthcare assist doctors and medical practitioners with accurate diagnosis [38, 171]. Criminal risk assessment tools operate with judges to provide an empirical estimate of recidivism risk [96, 127]. Human experts are also involved in auditing the outputs from automated models to detect errors for input samples where the automated system has insufficient experience, as observed in the case of child maltreatment hotline screening [65]. Many other examples of similar hybrid human-machine decision-making frameworks exist in literature [236, 282, 315].

For such *human-in-the-loop* frameworks, the approaches used for learning a classifier can often be different than those used in traditional supervised learning algorithms. Assuming one or more human experts are available to assist a classifier in

decision-making, an ideal training process should ensure that the capabilities and expertise of the humans are appropriately utilized to improve prediction accuracy or performance. However, since humans can have additional costs associated with their decisions (corresponding to time or resources invested to make predictions), the classifier will be expected to bear the primary decision-making load and humans should only be consulted when the classifier has low confidence in its decision. One can see that training human-in-the-loop frameworks can be more complex than traditional supervised learning; along with training an accurate classifier, the framework should also decipher the domains of expertise of different human experts so that they can be consulted appropriately. This field of research has seen a lot of recent interest due to the applicability of such frameworks in a variety of real-world settings. Algorithms to learn accurate human-in-the-loop frameworks have been forwarded by a number of recent studies [204, 218, 219, 253]. Chapter 6 proposes a novel learning algorithm for human-in-the-loop deferral frameworks, where the goal is to train a classifier that can either make an accurate decision or that defers the decision to an appropriate human expert when the classifier has low confidence in its decision. Considering that a number of applications are currently adopting automated decision-making systems, human-in-the-loop frameworks can allow such applications to smoothly and steadily transition from human decision-making to automated decision-making.

2.3 Social Biases in Automated Decision-Making

Either due to inappropriate data or due to imperfect model designs, automated decision-making frameworks currently display problematic social biases in their output. Applications where decision-making institutions have historically denied opportunities to the underprivileged groups of the population, e.g. credit lend-

ing [258], will still suffer from the impact of such *historical biases* when automation is incorporated into the decision-making framework. Years of discriminatory decision-making can corrupt the training datasets used to learn automated decision-making models. Corrupted datasets are indeed currently employed for creating models in many real-world applications, such as recruitment [78, 264], healthcare [235, 302], facial analysis [39, 269], risk assessment [13, 94], and predictive policing [272]. Furthermore, inappropriate processes for past and current data collection, aggregation, and processing of these datasets has compounded biases against minority groups. For example, survey instruments for data collection often use oversimplified race categorizations, which ignore the historical and political background that led to popular racial classifications [135]. Similarly, measurement errors in data collection can be disparately larger for the groups which have historically denied equal opportunities, leading to diminished information about individuals from the group [285]. Misrepresentation or under-representation of certain demographic groups in the data used to develop the decision-making model will affect the performance of the model for these marginalized groups. Inappropriate representation limits the amount of information that a trained model learns about the affected group and correspondingly results in larger errors when used for decision-making over this group [293].

Biases in data used for learning automated models can affect the outcome in many different problematic ways. When the model is used for resource allocation, as in the case of loan applications, admissions, risk assessment, or any other supervised learning application, biases in outcomes can result in disparate resource allocation across demographic groups, resulting in a denial of *equal opportunity* [20]. Representational biases can also affect public perceptions associated with misrepresented or under-represented groups. The negative portrayal of minority groups in the input data or the resulting decisions of automated frameworks propagate,

and sometimes even exacerbate, the negative stereotypes associated with these groups [166]. Beyond data biases, inappropriate model designs that do not account for the heterogeneity in the underlying population demographics can also result in disparate performance across groups [103].

In particular, inappropriate representation in a data collection can imply two different kinds of mismatches between the data collection and the underlying population. From a technical viewpoint, inappropriate representation can arise when the data collection inaccurately represents the underlying data distribution. For instance, the top Google Image Search results for the query "CEOs" contain around 11% images of women, while in reality, the percentage of women CEOs in the US is around 27% [50, 166]. In this case, the dataset (i.e., image search results) present an inaccurate depiction of reality; Chapter 4 presents detailed results demonstrating such biases in Google Image Search results for a variety of occupations and Chapter 5 provides evidence of dialect under-representation in automatically generated text summaries when using popular text summarization algorithms. While deviation from reality is one important kind of inappropriate representation, we might also consider a data collection to misrepresentative if it does not appropriately acknowledge all relevant demographic groups. Once again, consider the example of summarizing an image collection into a small subset. Suppose that the collection contains 100 images, with 50 images of white people, and 10 images each of Asian, Black, Hispanic, and Native American people. If the goal is to create a summary with just five images, it would be important to represent the diverse set of people in the summary by choosing one representative image for each ethnicity even though the ethnicity-distribution of the summary will not align with that of the original dataset. However, if we did create a summary whose ethnicity distribution is similar to the original dataset, then this summary will exclude images of people from at least one ethnicity, and fail to appropriately represent the un-

derlying dataset population. Differentiating between these two kinds of representational biases will become quite important when assessing methods to mitigate them in data collections and automatically-generated summaries, as we discuss in the next section and in other chapters of this dissertation.

For a detailed extensive survey of algorithmic biases and their impact, I recommend the following cited surveys [20, 208]. The above studies, nevertheless, provide clear evidence of the prevalence of social biases in the decisions of automated frameworks; correspondingly, it is important to design methods that do not let data or model biases affect the framework's decisions. The field of algorithmic fairness indeed aims to accomplish this goal.

2.4 Algorithmic Fairness

Algorithmic fairness methods or interventions attempt to correct social biases with respect to socially salient attributes or protected attributes in the outputs of learning models. By socially-salient attributes, I refer to presented demographic attributes like gender or race as well as perceived attributes like skintone, dialect, or perceived gender. Protected attributes are commonly used to refer to attributes that are protected by law and discrimination with respect to these attributes is considered unlawful [290]. However, throughout this thesis, I will use the terms socially-salient attributes and protected attributes interchangeably since the goal of the methods proposed in this is to protect groups defined by both presented and perceived minority demographic attributes against algorithmic harms.

Algorithmic interventions to address social biases are usually designed individually for different kinds of learning approaches but the general approach is to create holistic models that do not propagate the biases of the data or associated humans. The outputs of the *fair* models are expected to satisfy some form of *statistical*

fairness property, usually quantified using a *fairness metric*.

For unsupervised learning algorithms, various notions of statistical fairness can be employed to ensure that representational biases in the data do not affect the outcome of the algorithms. The goal of *fairness* in this setting is to ensure that all demographic groups are appropriately represented. The statistical fairness metric used here usually corresponds to achieving *equal representation* of all groups in the output (e.g., an equal number of images of men and women in the summary of an image dataset) or *proportional representation* of all groups in the output (e.g., the proportion of the number of images of men and women in the summary should be similar to their proportion in reality). The choice between equal representation and proportional representation (or something in between) crucially depends on the application in question and the kind of representational bias that the fairness intervention aims to fix. As mentioned in the previous section, representational bias can either correspond to deviation from underlying population distribution and/or failure to represent all demographic groups appropriately. The choice of fairness intervention and parameters will hence depend on the kind of bias being exhibited in the decisions of the automated unsupervised models.

Efficient approaches to achieve representation have been proposed for all kinds of unsupervised learning methods. For summarization or ranking, proposed fairness interventions include using group-specific scoring functions [193] or constrained optimization algorithms with representation constraints [52, 77]. For clustering, similar constrained optimization approaches can be employed to ensure relevant representation of all groups among cluster centers or within each cluster [17, 60, 61]. One of the issues that arise when employing these methods in practice is the unavailability of socially salient or protected attribute labels or group memberships of individual elements. Prior methods for fair summarization, in particular, rely on the availability of group information, and in its absence, implement-

ing these algorithms can be infeasible. Chapters 4 and 5 tackle this problem by suggesting algorithms for fair summarization that debias automatically-generated summaries using user-defined representative examples.

Fair supervised learning methods propose algorithmic interventions to learn classification models that provide similar predictive performance for all individuals independent of their protected attributes. In the context of data on humans, classification involves implementing an automated policy that can predict class labels corresponding to individuals for a specific task; for example, predicting the health risk score of a patient or predicting whether a loan application should be accepted or not. As mentioned in the previous section, a large amount of literature has pointed out social biases and negative stereotypes in training datasets. The classifiers trained using biased training datasets simulate an inaccurate relationship between individuals' attributes and class labels resulting in reduced performance for the groups that the dataset misrepresents. Even beyond training datasets, model misspecifications can negatively affect the performance of classifiers for disadvantaged groups [208]. One way of addressing social biases in supervised learning is to construct classifiers which have similar performance across all groups and which satisfy certain statistical group-fairness properties. Popular examples of desired fairness properties include statistical parity (equal selection rate across all groups defined by protected attributes) [55, 312], equalized odds (all groups defined by protected attributes should have equal group-specific false positive and true positive rates) [136], min-max fairness [88, 205] and many others [225, 299]. Papers in the field of fair supervised learning have indeed proposed a variety of algorithms to construct *fair classifiers* that satisfy (one or more) fairness properties [21]. In particular, there are three main types of approaches for fair classification: (a) pre-processing approaches that debias the training dataset by attempting to learn the underlying unbiased distribution of the dataset [42, 45, 56,

124, 301, 308, 314]; (b) in-processing approaches that propose optimization methods to learn a classifier while satisfy given fairness constraints [9, 55, 91, 98, 312, 316]; and (c) post-processing approaches that modify a trained classifier to ensure that the adjusted classifier has similar performance for all groups [16, 111, 136, 305]. Chapter 6 demonstrates the application of an in-processing approach to achieve statistical fairness using statistical parity and min-max fairness constraints.

Human-in-the-loop frameworks described earlier can address social biases when used in a manner that counteracts the automated decision-making framework and training data biases. However, if used inappropriately, they can exhibit additional biases arising from the prejudices of human experts. Human biases can be different than training data or model bias if the humans in the framework are different than the ones who labeled the datasets or the ones who developed the classification models [127, 208]. As such, it is important to develop methods for human-in-the-loop learning that can accurately determine the input subspaces where the relevant humans are inaccurate and/or biased; for instance, one can incorporate fairness constraints when training human-in-the-loop frameworks to regularize the learning algorithm to favor unbiased accurate experts for input. Indeed, the deferral training methods proposed in Chapter 6 provide methods to ensure that the outputs of the learned framework satisfy statistical fairness properties, such as demographic parity or min-max fairness.

Finally, even though the goal of algorithmic fairness methods is to achieve statistical fairness, a number of papers in this field do acknowledge that simply achieving statistical fairness may not always be not sufficient by itself and ensuring transparency, accountability, and community participation during deployment are important to completely utilize the effectiveness of fairness interventions [20, 69, 126, 169]. Community involvement – especially from historically-marginalized groups [222] – along with deeper evaluation of our current methods

of data collection and processing are necessary steps that have to be taken to improve trust in automated decision-making [292]. The methods proposed in this dissertation indeed encourage stakeholder participation. To attain the maximal impact of the proposed methods in addressing social biases, a broader analysis of the context and environment where these fairness interventions are implemented is always necessary. Incorporating the principles of transparency, accountability, and community participation along with novel algorithmic fairness interventions can allow for the development of robust automated frameworks and the research and discussions presented in this dissertation indeed aim to abide by them.

Chapter 3

Auditing for Diversity Using Representative Examples

Mechanisms to audit the *diversity* of a dataset are necessary to assess the shortcomings of the dataset in representing the underlying distribution accurately. In particular, any dataset containing information about people is expected to suitably represent all social groups (defined by attributes such as gender, race, or age) present in the underlying population in order to mitigate disparate outcomes and impacts in downstream applications [39, 44]. However, many real-world and popular data sources suffer from the problem of disproportionate representation of minority groups [232, 237]. For example, prior work has shown that the top results in Google Image Search for occupations are more gender-biased than the ground truth of the gender distribution in that occupation [50, 166, 278].

Given the existence of biased data collections in mainstream media and web sources, methods to audit the diversity of generic data collections can help quantify and assess the existing biases in multiple ways. First, it gives a baseline idea

This chapter is based on joint work with L. Elisa Celis and was published in the proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2021 [168]. I would like to thank Joy Buolamwini for providing access to the PPB dataset used in this chapter and Chapter 4.

of the demographic distribution in the collection and its deviation from the true distribution of the underlying population. Second, stereotypically-biased representation of a social group in any data collection can lead to further propagation of negative stereotypes associated with the group [68, 138, 306] and/or induce incorrect perceptions about the group [117, 275]. A concrete example is the evidence of stereotype-propagation via biased Google Search results [166, 232]. These stereotypes and biases can be further exacerbated via machine learning models trained on the biased collections [39, 44, 237]. Providing an easy way to audit the diversity in these collections can help the users of such collections assess the potential drawbacks and pitfalls of employing them for downstream applications.

Auditing the diversity of any collection with respect to a protected attribute primarily involves looking at the *disparity* or *imbalance* in the empirical marginal distribution of the collection with respect to the protected attribute. For example, from prior work [50], we know that the top 100 Google Image Search results for CEOs in 2019 contained around 89 images of men and 11 images of women; in this case, we can quantify the disparity in this dataset, with respect to gender, as the difference between the fraction of minority group images and the fraction of majority group images, i.e., as $0.11 - 0.89 = -0.78$. The sign points to the direction of the disparity while the absolute value quantifies the extent of the disparity in the collection. Now suppose that, instead of just 100 images, we had multiple collections with thousands of query-specific images, as in the case of Google Image Search. Since these images have been scraped or generated from different websites, the protected attributes of the people in the images will likely not be labeled at the source. In the absence of protected attribute information, the task of simply auditing the diversity of these large collections (as an end-user) becomes quite labor-intensive. Hand-labeling large collections can be extremely time-expensive while using crowd-annotation tools (e.g. Mechanical Turk) can be very costly. For

a single collection, labeling a small subset (sampled i.i.d. from the collection) can be a reasonable approach to approximate the disparity; however, for multiple collections, this method is still quite expensive since, for every new collection, we will have to re-sample and label a new subset. It also does not support the addition/removal of elements to the collection. One can, alternately, use automated models to infer the protected attributes; although, for most real-world applications, these supervised models need to be trained on large labeled datasets (which may not be available) and pre-trained models might encode their own pre-existing biases [39]. The question, therefore, arises *if there is a cost-effective method to audit the diversity of large collections from a domain when the protected attribute labels of elements in the collections are unknown* (stated formally in Section 3.2).

The primary contribution of this chapter is an algorithm to evaluate the diversity of a given unlabeled collection with respect to any protected attribute (Section 3.3). The proposed algorithm takes as input the collection to be audited, a small set of labeled representative elements, called the *control set*, and a metric that quantifies the similarity between any given pair of elements. Using the control set and the similarity metric, our algorithm returns a proxy score of disparity in the collection with respect to the protected attribute. The same control set can be used for auditing the diversity of any collection from the same domain.

The control set and the similarity metric are the two pillars of our algorithm, and we theoretically show the dependence of the effectiveness of our framework on these components. In particular, the proxy measure returned by our algorithm approximates the true disparity measure with high probability, with the approximation error depending on the size and quality of the control set, and the quality of the similarity metric. The protected attributes of the elements of the control set are expected to be labeled; however, the primary advantage of our algorithm is that the size of the control set can be much smaller than the size of the collection to

achieve a small approximation error (Section 3.4.2). Empirical evaluations on the Pilots Parliamentary Benchmark (PPB) dataset [39] show that our algorithm, using randomly chosen control sets and cosine similarity metric, can indeed provide a reasonable approximation of the underlying disparity in any given collection (Section 3.4.1).

To further reduce the approximation error, we propose an algorithm to construct *adaptive* control sets (Section 3.5). Given a small labeled auxiliary dataset, our proposed control set construction algorithm selects the elements that can best differentiate between samples with the same protected attribute type and samples with different protected attribute types. We further ensure that the elements in the chosen control set are *non-redundant* and *representative* of the underlying population. Simulations on the PPB dataset, CelebA dataset [199] and TwitterAAE dataset [29] show that using the cosine similarity metric and adaptive control sets, we can effectively approximate the disparity in random and topic-specific collections, with respect to a given protected attribute (Section 3.6).

3.1 Related Work

With rising awareness around the existence and harms of algorithmic biases, prior research has explored and quantified disparities in data collections from various domains. When the dataset in consideration has labeled protected attributes, the task of quantifying the disparity is relatively straightforward. For instance, Davidson et al. [81] demonstrate racial biases in automated offensive language detection by using datasets containing Twitter posts with dialects labeled by the authors or domain experts. Larrazabal et al. [181] can similarly analyze the impact of gender-biased medical imaging datasets since the demographic information associated with the images is available at the source. However, as mentioned earlier, pro-

tected attribute labels for elements in a collection may not be available, especially if the collection contains elements from different sources.

In the absence of protected attribute labels from the source, crowd-annotation is one way of obtaining these labels and auditing the dataset. To measure the gender disparity in Google Image Search results, Kay et al. [166] crowd-annotated a small subset of images and compared the gender distribution in this small subset to the true gender distribution in the underlying population. Other papers on diversity evaluation have likewise used a small labeled subset of elements [35, 271] to derive inferences about larger collections. As discussed earlier, the problem with this approach is that it assumes that the disparity in the small labeled subset is a good approximation of the disparity in the given collection. This assumption does not hold when we want to estimate the diversity of new or multiple collections from the same domain or when elements can be continuously added/removed from the collection. Our method, instead, uses a given small labeled subset to approximate the disparity measure of any collection from the same domain. Semi-supervised learning also explores learning methods that combine labeled and unlabeled samples [326]. The labeled samples are used to train an initial learning model and the unlabeled samples are then employed to improve the model generalizability. Our proposed algorithm has similarities with the semi-supervised self-training approach [18], but is faster and more cost-efficient (Section 3.4.2).

Representative examples have been used for other bias-mitigation purposes in recent literature, such as fair data generation [63]. Kallus et al. [158] also employ reference sets for bias assessments; they approximate the disparate impact of prediction models in the absence of protected attribute labels. In comparison, our goal is to evaluate representational biases in a given collection. Chapters 4 and 5 also use control or reference sets for gender and skintone-diverse image summarization and dialect-diverse text summarization respectively.

3.2 Notations

Let $S := \{x_j\}_{j=1}^{N}$ denote the collection to be evaluated. Each element in the collection consists of a d-dimensional feature vector x, from domain $\mathcal{X} \subseteq \mathbb{R}^d$. Every element j in S also has a protected attribute, $z_j \in \{0, 1\}$, associated with it; however, we will assume that the protected attributes of the elements in S are unknown. Let $S_i := \{x_j, j \in [N] \mid z_j = i\}$. A measure of disparity in S with respect to the protected attribute is $d(S) := |S_0|/|S| - |S_1|/|S|$, i.e., the difference between the fraction of elements from group 0 and group 1. A dataset S is considered to be *diverse* with respect to the protected attribute if this measure is 0, and high $|d(S)|$ implies low diversity in S. Our goal will be to estimate this value for any given collection [1] [2]. Let p_{data} denote the underlying distribution of the collection S.

Control Set. Let $T := \left\{x'_j, z'_j\right\}_{j=1}^{m}$ denote the control set of size m, i.e., a small set of representative examples. Every element T also has a feature vector from domain $\mathcal{X}$ and a protected attribute associated with it. Let $T_i := \left\{x'_j, j \in [m] \mid z_j = i\right\}$. Importantly, the protected attributes of the elements in the control set are known and we will primarily employ control sets that have an equal number of elements from both protected attribute groups, i.e., $|T_0| = |T_1|$. The size of the control set is also much smaller than the size of the collection being evaluated, i.e., $|T| \ll |S|$. Let $\mathrm{p}_{\text{control}}$ denote the underlying distribution of the control set T.

Throughout this chapter, we will also use the notation $a \in b \pm c$ to denote that $a \in [b - c, b + c]$. The problem we tackle in this chapter is auditing the diversity of S using T; it is formally stated below.

[1] Our proposed method can be used for other metrics that estimate imbalance in the distribution of protected attribute as well (such as $|S_0|/|S|$); however, for the sake of simplicity, we will limit our analysis to $d(S)$ evaluation.

[2] We present the model and analysis for binary protected attributes. To extend the framework for non-binary protected attributes with k possible values, one can alternately define disparity as $\max_{i \in [k]} |S_i| - \min_{i \in [k]} |S_i|$.

Problem 3.2.1. *Given a collection S (with **unknown** protected attributes of elements) and a balanced control set T (with **known** protected attributes of elements), can we use T to approximate $d(S)$?*

3.3 Model and Algorithm

The main idea behind using the control set T to solve Problem 3.2.1 is the following: for each element $x \in S$, we can use the partitions T_0, T_1 of the control set to check which partition is most *similar* to x. If most elements in S are *similar* to T_0, then S can be said to have more elements with protected attribute z=0 (similarly for z=1). However, to employ this audit mechanism we need certain conditions on the *relevance* of the control set T, as well as, a metric that can quantify the similarity of an element in S to control set partitions T_0, T_1. We tackle each issue independently below.

3.3.1 Domain-relevance of the control set

To ensure that the chosen control set is representative and relevant to the domain of the collection in question, we will need the following assumption.

Assumption 3.3.1. *For any $x \in \mathcal{X}$, $p_{data}(x|z) = p_{control}(x|z)$, for all $z \in \{0, 1\}$.*

This assumption states that the elements of the control set are from the same conditional distribution as the elements of the collection S. It roots out settings where one would try to use non-representative control sets for diversity audits (e.g., full-body images of people to audit the diversity of a collection of portrait images). Note that despite similar conditional distributions, the control set and the collection can (and most often will) have different protected attribute marginal distributions.

We will use the notation $p_z(x)$ to denote the conditional distribution of x given z in the rest of the document, i.e., $p_z(x) := \mathrm{p}_{\text{data}}(x \mid z) = \mathrm{p}_{\text{control}}(x \mid z)$. Given a collection S, we will call a control set T (with partitions T_0, T_1) *domain-relevant* if the underlying distribution of T satisfies Assumption 3.3.1.

3.3.2 Similarity metrics

Note that even though $p_z(x)$ is the same for both the control set and the collection, the distributions $p_0(x)$ and $p_1(x)$ can be very different from each other, and our aim we will be to design and use similarity metrics that can differentiate between elements from the two conditional distributions.

A general pairwise similarity matrix sim : $\mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ takes as input two elements and returns a non-negative score of similarity between the elements; the higher the score, the more similar are the elements. For our setting we need a similarity metric that can, *on average*, differentiate between elements that have the same protected attribute type and elements that have different protected attribute types. Formally, we define such a similarity metric as follows.

Definition 3.3.1 (γ-similarity metric)**.** *Suppose we are given a similarity metric sim* : $\mathcal{X} \times \mathcal{X} \to [0, 1]$*, such that*

$$\mathbb{E}_{x_1, x_2 \sim p_z}\left[sim(x_1, x_2)\right] = \mu_{same} \quad \textit{and} \quad \mathbb{E}_{x_1 \sim p_{z_1}, x_2 \sim p_{z_2}, z_1 \neq z_2}\left[sim(x_1, x_2)\right] = \mu_{diff}.$$

Then for $\gamma \geq 0$, we call sim a γ-similarity metric if $\mu_{same} - \mu_{diff} \geq \gamma$.

Note that the above definition is not very strict; we do not require $\mathrm{sim}(\cdot, \cdot)$ to return a large similarity score for every pair of elements with the same protected attribute type or to return a small similarity score for every pair of elements with different protected attribute types. Rather $\mathrm{sim}(\cdot, \cdot)$, only *in expectation*, should be able to differentiate between elements from the same groups and elements from different

groups. In a later section, we show that the cosine similarity metric indeed satisfies this condition for real-world datasets.

3.3.3 Algorithm

Suppose we are given a *domain-relevant* control set T that satisfies Assumption 3.3.1 (with partitions T_0 and T_1) and a γ-similarity metric $\text{sim}(\cdot,\cdot)$. With slight abuse of notation, for any element $x \in S$, let $\text{sim}(x, T_i) = {}^1/_{|T_i|} \sum_{x' \in T_i} \text{sim}(x, x')$ and let $\text{sim}(S, T_i) = {}^1/_{|S|} \sum_{x \in S} \text{sim}(x, T_i)$. Let $\hat{d}(S) := \text{sim}(S, T_0) - \text{sim}(S, T_1)$. We propose the use of $\hat{d}(S)$ (after appropriate normalization) as a proxy measure for $d(S)$; Algorithm 1 presents the complete details of this proxy diversity score computation and Section 3.3.4 provides bounds on the approximation error of $\hat{d}(S)$. We will refer to Algorithm 1 by *DivScore* for the rest of the chapter.

Algorithm 1 *DivScore*: Algorithm for diversity audit

Input: Dataset S, control set $T := T_0 \cup T_1$, similarity metric $\text{sim}(\cdot,\cdot)$
Output: Approximation of the disparity score $d(S)$

$l \leftarrow \frac{1}{|T_0| \cdot |T_1|} \sum_{x,x' \in T_0 \times T_1} \text{sim}(x, x')$
$u_0 \leftarrow \frac{1}{|T_0| \cdot (|T_0|-1)} \sum_{x \in T_0, x' \in T_0 \setminus \{x\}} \text{sim}(x, x')$
$u_1 \leftarrow \frac{1}{|T_1| \cdot (|T_1|-1)} \sum_{x \in T_1, x' \in T_1 \setminus \{x\}} \text{sim}(x, x')$
Compute $\text{sim}(S, T_0) \leftarrow \frac{1}{|S| \cdot |T_0|} \sum_{x,x' \in S \times T_0} \text{sim}(x, x')$
$s_0 \leftarrow \frac{\text{sim}(S,T_0) - l}{u_0 - l}$
Compute $\text{sim}(S, T_1) \leftarrow \frac{1}{|S| \cdot |T_1|} \sum_{x,x' \in S \times T_1} \text{sim}(x, x')$
$s_1 \leftarrow \frac{\text{sim}(S,T_1) - l}{u_1 - l}$
return $s_0 - s_1$

3.3.4 Theoretical analysis

To prove that $\hat{d}(S)$ is a good proxy measure for auditing diversity, we first show that if $x \in S_i$, then $\text{sim}(x, T_i) > \text{sim}(x, T_j)$, for $j = 1 - i$, with high probability

and quantify the exact difference using the following lemma. For the analysis in this section, assume that the elements in T_0, T_1 have been sampled i.i.d. from conditional distribution p_0, p_1 respectively and $|T_0| = |T_1|$.

Lemma 3.3.2. *For $i \in \{0,1\}$, any $x \in S_i$ and $\delta > 0$, with probability at least $1 - 2e^{-\delta^2 \mu_{diff}|T|/6} \cdot (1 + e^{-\delta^2 \gamma |T|/6})$, we have*

$$sim(x, T_i) - sim(x, T_{1-i}) \in \mu_{same} - \mu_{diff} \pm \delta(\mu_{same} + \mu_{diff}). \tag{1}$$

The lemma basically states that a γ-similarity metric, with high probability, can differentiate between $\text{sim}(x, T_i)$ and $\text{sim}(x, T_{1-i})$. The proof uses the fact that since T is domain-relevant and the elements of T are i.i.d. sampled from the conditional distributions, for any $x' \in T_0$, $\mathbb{E}[\text{sim}(x, x')] = \mu_{\text{same}}$ and for any $x' \in T_1$, $\mathbb{E}[\text{sim}(x, x')] = \mu_{\text{diff}}$. Then, the statement of the lemma can be proven using standard Chernoff-Hoeffding concentration inequalities [147, 215]. Note that even though $\text{sim}(\cdot, \cdot)$ was defined to differentiate between protected attribute groups in expectation, by averaging over all control set elements in T_0, T_1, we are able to differentiate across groups with high probability. The proof of the lemma is presented below.

Proof of Lemma 3.3.2. Suppose x has protected attribute type 0, i.e., $x \in S_0$. Since control set T is domain-relevant, we know that for any $x' \in T_0$, $\mathbb{E}[\text{sim}(x, x')] = \mu_{\text{same}}$ and for any $x' \in T_1$, $\mathbb{E}[\text{sim}(x, x')] = \mu_{\text{diff}}$. Then, using Chernoff-Hoeffding bounds [147, 215], we get that for any $\delta > 0$,

$$\mathbb{P}\left[\text{sim}(x, T_0) \notin (1 \pm \delta)\mu_{\text{same}}\right] \leq 2\exp\left(-\delta^2 \cdot |T_0| \cdot \mu_{\text{same}}/3\right), \text{ and}$$

$$\mathbb{P}\left[\text{sim}(x, T_1) \notin (1 \pm \delta)\mu_{\text{diff}}\right] \leq 2\exp\left(-\delta^2 \cdot |T_1| \cdot \mu_{\text{diff}}/3\right).$$

Note that $|T_0| = |T_1| = |T|/2$. The probability that both the above events are

simultaneously satisfied is

$$2\exp\left(-\delta^2\mu_{\text{same}}\frac{|T|}{6}\right)+2\exp\left(-\delta^2\mu_{\text{diff}}\frac{|T|}{6}\right) \\ \leq 2\exp\left(-\delta^2\mu_{\text{diff}}\frac{|T|}{6}\right)\cdot\left(1+\exp\left(-\delta^2\gamma\frac{|T|}{6}\right)\right).$$

Therefore, combining the two statements we get that with probability at least $1-2\exp\left(-\delta^2\mu_{\text{diff}}|T|/6\right)\cdot\left(1+\exp\left(-\delta^2\gamma|T|/6\right)\right)$,

$$\text{sim}(x,T_0)-\text{sim}(x,T_1)\in\left[(1-\delta)\mu_{\text{same}}-(1+\delta)\mu_{\text{diff}},(1+\delta)\mu_{\text{same}}-(1-\delta)\mu_{\text{diff}}\right].$$

Simplifying the above expression, we get

$$\text{sim}(x,T_i)-\text{sim}(x,T_{1-i})\in\mu_{\text{same}}-\mu_{\text{diff}}\pm\delta(\mu_{\text{same}}+\mu_{\text{diff}}).$$

The other direction (when $x\in S_1$) follows from symmetry. □

Lemma 3.3.2 also partially quantifies the dependence on $|T|$ and γ. Increasing the size of the control set T will lead to a higher success probability. Similarly, larger γ implies that the similarity metric is more powerful in differentiating between the groups, which also leads to a higher success probability. Using the above lemma, we can next prove that the proposed diversity audit measure is indeed a good approximation of the disparity in S. Recall that, for the dataset S, $\text{sim}(S,T_i)=\frac{1}{|S|}\sum_{x\in S}\text{sim}(x,T_i)$.

Theorem 3.3.3 (Diversity audit measure)**.** *For protected attribute $z\in\{0,1\}$, let p_z denote the underlying conditional distribution $p_{data}(x|z)$. Suppose we are a given a dataset S containing i.i.d. samples from p_{data}, a domain-relevant control set T (with pre-defined partitions by protected attribute T_0 and T_1, such that $|T_0|=|T_1|$) and a similarity metric sim : $\mathcal{X}^2\to\mathbb{R}_{\geq 0}$, such that if $\mu_{same}=\mathbb{E}_{x_0,x_1\sim p_z}\left[sim(x_0,x_1)\right]$, $\mu_{diff}=\mathbb{E}_{x_0\sim p_0,x_1\sim p_1}\left[sim(x_0,x_1)\right]$,*

then $\mu_{same} - \mu_{diff} \geq \gamma$, *for* $\gamma > 0$. *Let*

$$\delta = \sqrt{\frac{6\log(20|S|)}{|T|\min(\mu_{diff}, \gamma)}}$$

and let $\hat{d}(S) := sim(S, T_0) - sim(S, T_1)$. *Then, with high probability,* $\hat{d}(S)/(\mu_{same} - \mu_{diff})$ *approximates* $d(S)$ *within an additive error of* $\delta \cdot (\mu_{same} + \mu_{diff})/(\mu_{same} - \mu_{diff})$. *In particular, with probability* $\gtrapprox 0.9$,

$$\hat{d}(S) \in (\mu_{same} - \mu_{diff}) \cdot d(S) \pm \delta \cdot (\mu_{same} + \mu_{diff}).$$

Proof of Theorem 3.3.3. Applying Lemma 3.3.2 to each element in S, we get that with probability at least $q := 1 - 2|S|e^{-\delta^2\mu_{\text{diff}}|T|/6} \cdot (1 + e^{-\delta^2\gamma|T|/6})$, all elements satisfy condition (1). Summing up $\text{sim}(x, T_0) - \text{sim}(x, T_1)$ for all $x \in S$, we get

$$\text{sim}(S, T_0) - \text{sim}(S, T_1) \in (\mu_{\text{same}} - \mu_{\text{diff}}) \cdot \frac{|S_0| - |S_1|}{|S|} \pm \delta(\mu_{\text{same}} + \mu_{\text{diff}}).$$

Simplifying the above bound, we have that with probability q,

$$\hat{d}(S) \in (\mu_{\text{same}} - \mu_{\text{diff}}) \cdot d(S) \pm \delta \cdot (\mu_{\text{same}} + \mu_{\text{diff}}).$$

By choosing $\delta = \sqrt{\frac{6\log(20|S|)}{|T|\min(\mu_{\text{diff}}, \lambda)}}$, the probability q is at least

$$\begin{aligned} 1 - 2|S|e^{-\delta^2\mu_{\text{diff}}|T|/6}(1 + e^{-\delta^2\gamma|T|/6}) &\geq 1 - 2|S|e^{-\log 20|S|}(1 + e^{-\log 20|S|}) \\ &= 0.9 - \frac{1}{200|S|}. \end{aligned}$$

□

Theorem 3.3.3 basically states that, with high probability, $d(S)$ is contained in a

small range of values determined by $\hat{d}(S)$, i.e.,

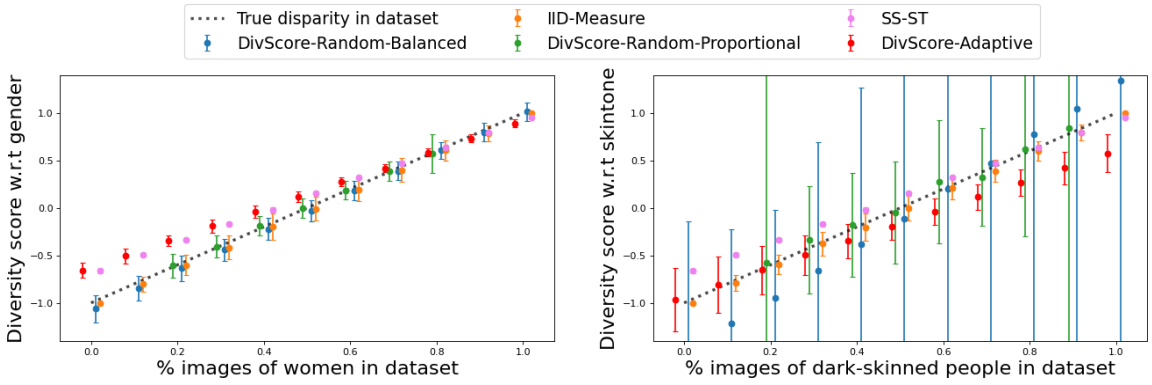
$$d(S) \in \left(\hat{d}(S) \pm \delta \cdot (\mu_{\text{same}} + \mu_{\text{diff}})\right) / (\mu_{\text{same}} - \mu_{\text{diff}})) .$$

The theoretical analysis is in line with the implementation in Algorithm 1 (*DivScore*), i.e., the algorithm computes $\hat{d}(S)$ and normalizes it appropriately using estimates of μ_{same} and μ_{diff} derived from the control set.

Note that Theorem 3.3.3 assumes that $\mu_{\text{same}} = \mathbb{E}_{x_0,x_1\sim p_z}[\text{sim}(x_0, x_1)]$ is the same for both $z \in \{0, 1\}$. However, they may not be the same in practice and *DivScore* uses separate upper bounds for z=0 and z=1 (u_0 and u_1 respectively). Similarly, we don't necessarily require a balanced control set (although, as discussed in Section 3.4.2, a balanced control set is preferable over an imbalanced one). We keep the theoretical analysis simple for clarity, but both these changes can be incorporated in Theorem 3.3.3 to derive similar bounds as well.

The dependence of error on γ and T can also be inferred from Theorem 3.3.3. The denominator in the error term in Theorem 3.3.3 is lower bounded by γ. Therefore, the larger the γ, the lower the error and the tighter the bound. The theorem also gives us an idea of the size of the control set required to achieve low δ error and high success probability. To keep δ small, we can choose a control set T with $|T| = \Omega(\log |S|)$. In other words, a control set of size $c \log |S|$ elements, for an appropriate $c > 1$, should be sufficient to obtain a low approximation error. Since the control sets are expected to have protected attribute labels (to construct partitions T_0 and T_1), having small control sets will make the usage of our audit algorithm much more tractable.

Cost of *DivScore*. The time complexity of Algorithm 1 (*DivScore*) is $O(|S|\cdot|T|)$, and it only requires $|T|$ samples (control set) to be labeled. In comparison, if one was to label the entire collection to derive $d(S)$, the time complexity would be



(a) Gender protected attribute

(b) Skin-tone protected attribute

Figure 3.1: Results for PPB-2017 dataset using random and adaptive control sets. The reported performance is the mean of output from *DivScore* across 100 repetitions (error bars denote standard error). To improve readability, we limit the y-axis to the range to $[-1.5, 1.5]$, which results in trimmed errorbands for some methods; we present the same expanded plots without axis restrictions in Appendix A.1.2. The protected attributes considered here are gender and skintone. The x-axis reports the fraction of $z = 0$ images in the collection (with set $\{0, 0.1, 0.2, \ldots, 1.0\}$ as the range of values) and, for each collection, we report the following five metrics in the y-axis: true disparity of the collection, *DivScore-Random-Balanced*, *DivScore-Random-Proportional*, *IID-Measure*, and *DivScore-Adaptive*. A collection is considered diverse if the diversity score (y-axis) is 0; the larger the deviation of the diversity score from 0, the lower the diversity is in the evaluated collection. Amongst all metrics, *DivScore-Adaptive*, *IID-Measure*, and *SS-ST* seem to have the lowest standard error. However, using *IID-Measure* and *SS-ST* are much costlier than *DivScore*, as discussed in Section 3.4.2.

$O(|S|)$, but all $|S|$ samples would need to be labeled. With a control set T of size $\Omega(\log |S|)$, our approach is much more cost-effective. The elements of T are also not dependent on elements of S; hence, the same control set can be used for other collections from the same domain.

3.4 Empirical Evaluation Using Random Control Sets

We first demonstrate the efficacy of the *DivScore* algorithm on a real-world dataset using random, *domain-relevant* control sets.

3.4.1 PPB-2017 dataset

The PPB (Pilots Parliamentary Benchmark) dataset consists of 1270 portrait images of parliamentarians from six different countries[3]. The images in this dataset are labeled with gender (male vs female) and skintone (values are the 6 types from the Fitzpatrick skin-type scale [112]) of the person in the image. This dataset was constructed and curated by Buolamwini and Gebru [39]. We will use gender and skintone as the protected attributes for our diversity audit analysis.

Methodology. We first split the dataset into two parts: the first containing 200 images and the second containing 1070 images. The first partition is used to construct control sets, while the second partition is used for diversity audit evaluation. Since we have the gender and skin-tone labels for all images, we can construct sub-datasets of size 500 with a custom distribution of protected attribute types. In other words, for a given $f \in \{0, 0.1, 0.2, \ldots, 1.0\}$, we construct a sub-dataset S of the second partition containing $f \cdot |S|$ images corresponding to protected attribute $z = 0$. Hence, by applying Algorithm 1 (*DivScore*) using a given control set T, we can assess the performance of our proxy measure for collection with a varying fraction of under/over-represented group elements.

When the protected attribute is gender, $z = 0$ will denote $g =$ female, when the protected attribute is skintone, $z = 0$ will denote $s > 3$ (skin-tone types corresponding to dark skin), and when the protected attribute is the intersection of gender and skin-tone, $z = 0$ will denote $g =$ female and $s > 3$ (corresponding to dark-skinned women).

Control sets. To evaluate the performance of *DivScore* the selection of elements for the control sets (of size 50 from the first partition) can be done in multiple

[3] `gendershades.org`

ways: (1) *random balanced control sets*, i.e., randomly block-sampled control sets with an equal number of $z = 0$ and $z = 1$ images; (2) *random proportional control sets*, i.e., control sets sampled i.i.d. from the collection in question; (3) *adaptive control sets*, i.e., non-redundant control sets that can best differentiate between samples with the same protected attribute type and samples with different protected attribute types. The complete details of the construction of *adaptive control sets* are given in Section 3.5; in this section, we primarily focus on the performance of *DivScore* when using random control sets. We will refer to our method as *DivScore-Random-Balanced*, when using random balanced control sets, and as *DivScore-Random-Proportional*, when using random proportional control sets. In expectation, random proportional control sets will have a similar empirical marginal distribution of protected attribute types as the collection; correspondingly, we also report the disparity measure of the random proportional control set $d(T)$ as a baseline. We will refer to this baseline as *IID-Measure*. Random proportional control sets need to be separately constructed for each new collection, while the same random balanced control set can be used for all collections; we discuss this contrast further in Section 3.4.2.

We also implement a semi-supervised self-training algorithm as a baseline. This algorithm (described formally in Appendix A.1.1) iteratively labels the protected attribute of those elements in the dataset for which the similarity to one group in the control set is significantly larger than the similarity to the other group. It then uses the learned labels to compute the diversity score. We implement this baseline using random control sets and refer to it as *SS-ST*. [4]

[4]We do not compare against crowd-annotation since the papers providing crowd-annotated datasets in our considered setting usually do not have ground truth available to estimate the approximation error.

Similarity Metric. We construct feature vector representations for all images in the dataset using pre-trained deep image networks. The feature extraction details are presented in Appendix A.1.1. Given the feature vectors, we use the cosine similarity metric to compute the pairwise similarity between images. In particular, given feature vectors x_1, x_2 corresponding to any two images, we will define the similarity between the elements as

$$\text{sim}(x_1, x_2) := 1 + \frac{x_1^\top x_2}{\|x_1\|\|x_2\|}. \tag{2}$$

We add 1 to the standard cosine between two vectors to ensure that the similarity values are always non-negative.

Evaluation Measures. We repeat the simulation 100 times; for each repetition, we construct a new split of the dataset and sample a new control set. We report the true fraction f and the mean (and standard error) of all metrics across all repetitions.

Results. The results are presented in Figure 3.1 (the figure also plots the performance of *DivScore-Adaptive*, which is discussed in Section 3.5). With respect to gender, Figure 3.1a shows that the *DivScore* measure is always close to the true disparity measure for all collections, and the standard error of all metrics is quite low. In this case, random control sets (balanced or proportional) can indeed approximate the disparity of all collections with very small errors.

The results are more mixed when skintone is the protected attribute. Figure 3.1b shows that while the *DivScore* average is close to the true disparity measure, the standard errors are quite high. The baselines *IID-Measure* and *SS-ST* have lower errors than our proxy measure (although they are not feasible methods for real-world applications, as discussed in the next section). The poor performance for this protected attribute, when using random control sets, suggests that strategies

to construct *good* non-random control sets are necessary to reduce the approximation error.

3.4.2 Discussion

The presented algorithm, *DivScore*, seems simple and efficient at first glance. While simplicity is indeed a feature of this algorithm, the efficiency depends on a variety of components. In this section, we discuss how different choices for these components control the efficiency of *DivScore*.

Dependence on γ. The performance of *DivScore* on PPB-dataset highlights the dependence of approximation error on the γ. Since the gender and skintone labels of images in the dataset are available, we can empirically derive the γ value for each protected attribute using the cosine similarity metric. When gender is the protected attribute, γ is around 0.35. On the other hand, when skintone is the protected attribute, γ is 0.08. In other words, the cosine similarity metric is able to differentiate between images of men and women to a better extent than between images of dark-skinned and light-skinned people. This difference in γ is the reason for the relatively larger error of *DivScore* in the case of skintone protected attribute.

Cosine similarity metric. The simulations also show that measuring similarity between images using the cosine similarity metric over feature vectors from pre-trained networks is indeed a reasonable strategy for disparity measurement. Pre-trained image networks and cosine similarity metric have similarly also been used in prior work for classification and clustering purposes [229, 307]. Intuitively, the cosine similarity metric is effective when conditional distributions p_0 and p_1 are concentrated over separate clusters over the feature space; e.g., for PPB dataset and gender as the protected attribute, the high value of γ (0.35) provides evidence

of this phenomenon. In this case, cosine similarity can, *on average*, differentiate between elements from the same cluster and different clusters.

Dependence on $|T|$. The size of the control set is another factor that is inversely related to the error of the proxy disparity measure. For this section, we use control sets of size 50. Smaller control sets lead to larger variance, as seen in Figure A.2 in the Appendix, while using larger control sets might be inhibitory and expensive since, in real-world applications, protected attributes of the control set images need to be hand-labeled or crowd-annotated.

Nevertheless, these empirical results highlight the crucial dependence on γ and properties of the control set T. In the next section, we improve upon the performance of our disparity measure and reduce the approximation error by designing non-random control sets that can better differentiate across the protected attribute types.

Drawbacks of *IID-Measure*. Recall that *IID-Measure* essentially uniformly samples a small subset of elements of the collection and reports the disparity of this small subset. Figure 3.1 shows that this baseline indeed performs well for the PPB dataset. However, it is not a cost-effective approach for real-world disparity audit applications. The main drawback of this baseline is that the subset has to have i.i.d. elements from the collection being audited for it to accurately predict the disparity of the collection. This implies that, for every new collection, we will have to re-sample and label a small subset to audit its diversity using *IID-Measure*. It is unreasonable to apply this approach when there are multiple collections (from the same domain) that need to be audited or when elements are continuously being added/removed from the collection. The same reasoning limits the applicability of *DivScore-Random-Proportional*.

DivScore-Random-Balanced, on the other hand, addresses this drawback by us-

ing a generic labeled control set that can be used for any collection from the same domain, without the additional overhead of constructing a new control set every time. This is also why balanced control sets should be preferred over imbalanced control sets since a balanced control set will be more adept at handling collections with varying protected attribute marginal distributions.

Drawbacks of *SS-ST*. The semi-supervised learning baseline *SS-ST* has larger estimation bias than *DivScore-Random-Balanced* and *DivScore-Random-Proportional*, but has lower approximation error than these methods. However, the main drawback of this baseline is the time complexity. Since it iteratively labels elements and then adds them to the control set to use for future iterations, the time complexity of this baseline is quadratic in dataset size. In comparison, the time complexity of *DivScore* is linear in the dataset size.

3.5 Adaptive Control Sets

As mentioned in the above discussion, the performance of *DivScore* depends crucially on the choice of the control set. In this section, we present a method to find a control set using which *DivScore* can achieve a small approximation error.

The theoretical analysis in Section 3.3.4 and the simulations in Section 3.4.1 use random control sets; i.e., T contains i.i.d. samples from p_0 and p_1 conditional distributions. This choice was partly necessary because the error depends on the γ-value of the similarity metric, which is quantified as $\mu_{\text{same}} - \mu_{\text{diff}}$, where

$$\mu_{\text{same}} = \mathbb{E}_{x_0, x_1 \sim p_z} \left[\text{sim}(x_0, x_1) \right], \quad \mu_{\text{diff}} = \mathbb{E}_{x_0 \sim p_0, x_1 \sim p_1} \left[\text{sim}(x_0, x_1) \right].$$

However, quantifying $\mu_{\text{same}}, \mu_{\text{diff}}$ (and, hence, γ) using expectation over the entire distribution might be unnecessary. In particular, the theoretical analysis uses

μ_{same} to quantify $\mathbb{E}_{x\sim p_i}[\text{sim}(x, T_i)]$, for any $i \in \{0,1\}$ (similarly μ_{diff}). Hence, we require the difference between μ_{same} and μ_{diff} to be large only when comparing the elements from the underlying distribution to the elements in the control set. This simple insight provides us a way to choose *good* control sets; i.e., we can choose control sets T for which the difference $|\mathbb{E}_x[\text{sim}(x, T_i)] - \mathbb{E}_x[\text{sim}(x, T_{1-i})]|$ is large.

Control sets that maximize γ. Suppose we have an auxiliary set U of i.i.d. samples from p_{data}, such that the protected attributes of elements in U are known. Let U_0, U_1 denote the partitions with respect to the protected attribute. Once again, $U \ll |S|$ and U will be used to construct a control set T. Let $m \in \{0, 2, 4, \ldots, |U|\}$ denote the desired size of T. For each $i \in \{0,1\}$ and $y \in U_i$, we can first compute

$$\gamma_i^{(y)} := \mathbb{E}_{x\sim U_i\setminus\{y\}}[\text{sim}(x,y)] - \mathbb{E}_{x\sim U_{1-i}\setminus\{y\}}[\text{sim}(x,y)],$$

and then construct a control set T by adding $m/2$ elements from each U_i with the largest values in the set $\left\{\gamma_i^{(y)}\right\}_{y\in U_i}$ to T.

Reducing redundancy in control sets. While the above methodology will result in control sets that maximize the difference between similarity with same group elements vs similarity with different group elements, it can also lead to *redundancy* in the control set. For instance, if two elements in U are very similar to each other, they will large pairwise similarity and can, therefore, both have large $\gamma_i^{(y)}$ value ; however, adding both to the control set is redundant. Instead, we should aim to make the control set as *diverse* and *representative* of the underlying population as possible. To that end, we employ a Maximal Marginal Relevance (MMR)-type approach and iteratively add elements from U to the control set T. For the first $m/2$ iterations, we add elements from U_0 to T. Given a hyper-parameter $\alpha \geq 0$, at any iteration t, the element added to T is the one that maximizes the following

score:

$$\left\{\gamma_0^{(y)} - \alpha \cdot \max_{x \in T} \text{sim}(x, y)\right\}_{y \in U_0 \setminus T}.$$

The next $m/2$ iterations similarly adds elements from U_1 to T using $\gamma_1^{(y)}$. The quantity $\max_{x \in T} \text{sim}(x, y)$ is the *redundancy score* of y; i.e., the maximum similarity of y with any element already added to T. By penalizing an element for being very similar to an existing element in T, we can ensure that chosen set T is diverse. The complete algorithm to construct such a control set, using a given U, is provided in Algorithm 2. We will refer to the control sets constructed using Algorithm 2 as *adaptive* control sets and Algorithm 1 with adaptive control sets as *DivScore-Adaptive*.

Note that, even with this control set construction method, the theoretical analysis does not change. Given any control set T ($= T_0 \cup T_1$), let

$$\gamma^{(T)} := \mathbb{E}_i \left[\mathbb{E}_{x \in p_i} \left[\text{sim}(x, T_i)\right] - \mathbb{E}_{x \sim p_{1-i}} \left[\text{sim}(x, T_{1-i})\right]\right].$$

For a control set T with parameter $\gamma^{(T)}$, we can obtain the high probability bound in Theorem 3.3.3 by simply replacing γ by $\gamma^{(T)}$. In fact, since we are explicitly choosing elements that have large $\gamma_i^{(\cdot)}$ parameters, $\gamma^{(T)}$ is expected to be larger than γ and, hence, using the adaptive control set will lead to a stronger bound in Theorem 3.3.3.

Our algorithm uses the standard MMR framework to reduce redundancy in the control set. Importantly, prior work has shown that the greedy approach of selecting the *best available* element is indeed approximately optimal [46]. Other non-redundancy approaches, e.g., Determinantal Point Processes [177], can also be employed.

Algorithm 2 Algorithm to construct an *adaptive* control set

Input: Auxiliary set $U = U_0 \cup U_1$, similarity metric sim, m, $\alpha \geq 0$
Output: Control set T

$T_0, T_1, \gamma_0, \gamma_1 \leftarrow \emptyset$
for $i \in \{0, 1\}$ **do**
    for $x \in U_i$ **do**
        $\gamma_i^{(x)} \leftarrow \frac{1}{|U_i|-1} \sum_{y \in U_i \setminus \{x\}} \text{sim}(x, y) - \frac{1}{|U_{1-i}|} \sum_{y \in U_{1-i}} \text{sim}(x, y)$
    while $|T_i| < m/2$ **do**
        $T_i \leftarrow T_i \cup \left\{ \arg\max \left\{ \gamma_i^{(x)} - \alpha \cdot \max_{y \in T_i} \text{sim}(x, y) \right\}_{x \in U_i \setminus T_i} \right\}$
return $T_0 \cup T_1$

Cost of each method. *DivScore-Adaptive* requires an auxiliary labeled set U from which we extract a good control set. Since $|U| > |T|$, the cost (in terms of time and labeling required) of using *DivScore-Adaptive* is slightly larger than the cost of using *DivScore-Random-Balanced*, for which we just need to randomly sample $|T|$ elements to get a control set. However, results in Appendix A.1.2 show that to achieve a similar approximation error, the required size of adaptive control sets is smaller than the size of random control sets. Hence, even though adaptive control sets are more costly to construct, *DivScore-Adaptive* is more cost-effective for disparity evaluations and requires smaller control sets (compared to *DivScore-Random-Balanced*) to approximate with low error.

3.6 Empirical Evaluation using Adaptive Control Sets

3.6.1 PPB-2017

Once again, we first test the performance of adaptive control sets on the PPB-2017 dataset. Recall that we split the dataset into two parts of sizes 200 and 1070 each. Here, the first partition serves as the auxiliary set U for Algorithm 2. The input

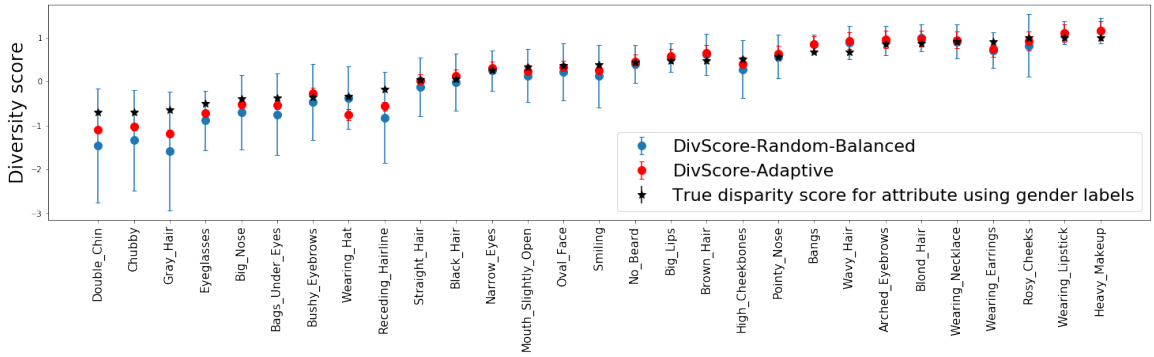


Figure 3.2: Results for CelebA dataset. For each feature, we plot the true gender disparity score for that feature as well as the scores obtained using *DivScore-Random-Balanced* and *DivScore-Adaptive* approaches. For both methods, the control set size is kept to 50. Note that the error of *DivScore-Adaptive* is much smaller in this case.

hyper-parameter α is set to be 1. The rest of the setup is the same as in Section 3.4.1.

Results. The results for this simulation are presented in Figure 3.1 (in red). The plots show that by using adaptive control sets, we obtain sharper proxy diversity measures for both gender and skintone. For skintone protected attribute, the standard error of *DivScore-Adaptive* is significantly lower than *DivScore-Random-Balanced*.

Note that the average of *DivScore-Adaptive*, across repetitions, do not align with the true disparity measure (unlike the results in the case of random control sets). This is because the adaptive control sets do not necessarily represent a uniformly random sample from the underlying conditional distributions. Rather, they are the subset of images from U with the best scope of differentiating between images from different protected attribute types. This non-random construction of the control sets leads to a possibly-biased but tighter approximation for the true disparity in the collection.

As noted before, when using adaptive control sets (from Algorithm 2), the performance depends on $\gamma^{(T)} := \mathbb{E}_i \left[\mathbb{E}_{x \in p_i} [\text{sim}(x, T_i)] - \mathbb{E}_{x \sim p_{1-i}} [\text{sim}(x, T_{1-i})] \right]$. By construction, we want to choose control sets T for which $\gamma^{(T)}$ is greater than the γ

value over the entire distribution. Indeed, in the case of the PPB dataset and for every protected attribute, we observe that $\gamma^{(T)}$ values of the adaptive control sets are much larger than the corresponding value when of randomly chosen control sets. When gender is the protected attribute, on average, $\gamma^{(T)}$ is 0.96 (for random control sets, it was 0.35). Similarly, when skintone is the protected attribute, $\gamma^{(T)}$ is around 0.34 (for random control sets, it was 0.08). The stark improvement in these values, compared to random control sets, is the reason behind the increased effectiveness of adaptive control sets in approximating the disparity of the collection.

3.6.2 CelebA dataset

CelebA dataset [199] contains images of celebrities with tagged facial attributes, such as whether the person in the image has eyeglasses, mustache, etc., along with the gender of the person in the image[5]. We use 29 of these attributes and a random subset of around 20k images for our evaluation. The goal is to approximate the disparity in the collection of images corresponding to a given facial attribute.

Methodology. We evaluate the performance of methods *DivScore-Random-Balanced* and *DivScore-Adaptive* for this dataset[6]. We perform 25 repetitions; in each repetition, an auxiliary set U is sampled of size 500 (and removed from the main dataset) and used to construct either a random control set (of size 50) or an adaptive control set (of size 50). The chosen control set is kept to be the same for all attribute-specific collections in a repetition. For each image, we use the pre-trained image networks to extract feature vectors (see Appendix A.1.1 for details) and the cosine similarity metric - Equation (2) - to compute pairwise similarity.

[5] `mmlab.ie.cuhk.edu.hk/projects/CelebA.html`

[6] For CelebA and TwitterAAE datasets, we only report the performance of *DivScore-Adaptive* and *DivScore-Random-Balanced* to ensure that the plots are easily readable. The performance of *DivScore-Random-Balanced* is similar to that of *DivScore-Random-Balanced* and, due to large data collection sizes, *SS-ST* is infeasible in this setting.

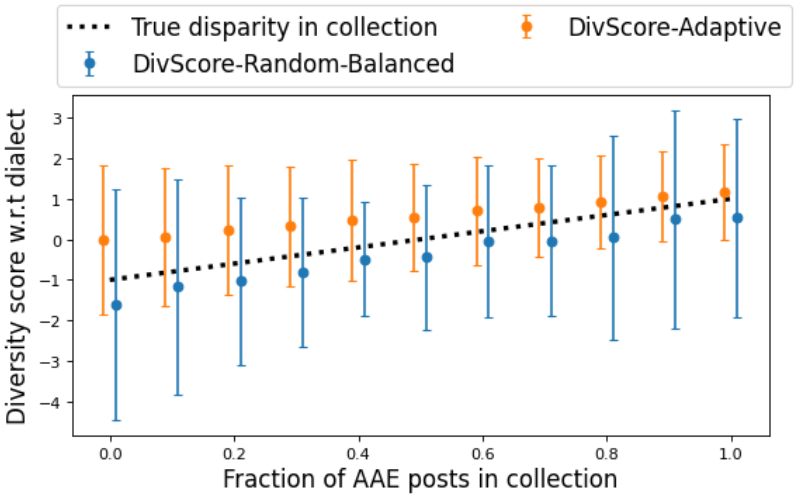


Figure 3.3: Results for TwittterAAE dataset with dialect as the protected attribute for *DivScore-Random-Balanced* and *DivScore-Adaptive* using control sets of size 50.

Results. The results are presented in Figure 3.2. The plot shows that, for almost all attributes, the score returned by *DivScore-Adaptive* is close to the true disparity score and has a smaller error than *DivScore-Random-Balanced*. Unlike the collections analyzed in PPB evaluation, the attribute-specific collections of the CelebA dataset are non-random; i.e., they are not i.i.d. samples from the underlying distribution. Nevertheless, *DivScore-Adaptive* is able to approximate the true disparity for each attribute-specific collection quite accurately.

Note that, for these attribute-specific collections, implementing *IID-Measure* would be very expensive, since one would have to sample a small set of elements for each attribute and label them. In comparison, our approach uses the same control set for all attributes and, hence, is much more cost-effective.

3.6.3 TwitterAAE dataset

To show the effectiveness of *DivScore* beyond image datasets, we analyze the performance over a dataset of Twitter posts. The *TwitterAAE* dataset, constructed by Blodgett et al. [29], contains around 60 million Twitter posts[7]. We filter the dataset

[7] `slanglab.cs.umass.edu/TwitterAAE/`

to contain only posts that either are *certainly* written in the African-American English (AAE) dialect (100k posts) or the White English dialect (WHE) (1.06 million posts). The details of filtering and feature extraction using a pre-trained Word2Vec model [210] are given in Appendix A.1.1.

Methodology. For this dataset, we will evaluate the performance of *DivScore-Random-Balanced* and *DivScore-Adaptive*[6]. We partition the datasets into two parts: the first contains 200 posts and the second contains the rest. The first partition is used to construct control sets of size 50 (randomly chosen from the first partition for *DivScore-Random-Balanced* and using Algorithm 2 for *DivScore-Adaptive*). The protected attribute is the dialect of the post. The second partition is used for diversity audit evaluation. We construct sub-datasets or collections with a custom distribution of posts from each dialect. For a given $f \in \{0, 0.1, \ldots, 1.0\}$, we construct a sub-dataset S of the second partition containing $f \cdot |S|$ AAE posts. The overall size of the sampled collection is kept to 1000 and we perform 25 repetitions. For *DivScore-Adaptive*, we use $\alpha = 0.1$.

Results. The audit results for collections from the TwitterAAE dataset are presented in Figure 3.3. The plot shows that both *DivScore-Random-Balanced* and *DivScore-Adaptive* can, on expectation, approximate the disparity for all collections; the disparity estimate from both methods increases with increasing fraction of AAE posts in the collection. However, once again, the approximation error of *DivScore-Adaptive* is smaller than the approximation error of *DivScore-Random-Balanced* in most cases[8].

[8]The code for this chapter is available at `https://github.com/vijaykeswani/Diversity-Audit-Using-Representative-Examples`.

3.7 Discussion, Limitations, and Future Work

As with any algorithm that aims to statistically model a real-world societal problem, there are questions about how generalizable are the results of the proposed algorithm. In this section, we discuss these questions, stating the potential applications of our framework, along with the practical limitations and directions for future work on real-world bias audits.

Third-party implementations and auditing summaries. To audit the diversity of any collection, *DivScore* simply requires access to a small labeled control set and a similarity metric. The cost of constructing these components is relatively small (compared to labeling the entire collection) and, hence, our audit framework can be potentially employed by third-party agencies that audit independently of the organization owning/providing the collections. For instance, our algorithm can be implemented as a browser plugin to audit the gender diversity of Google Image results or the dialect diversity of Twitter search results. Such a domain-generic diversity audit mechanism can be used to ensure a more-balanced power dynamic between the organizations disseminating/controlling the data and the users of the applications that use this data.

Variable-sized collections. *DivScore* can easily adapt to updates to the collections being audited. If an element is added/removed, one simply needs to add/remove the contribution of this element from $\text{sim}(S, T_0)$ and $\text{sim}(S, T_1)$, and recompute $\hat{d}(S)$. This feature crucially addresses the main drawback of *IID-Measure*.

Possibility of stereotype exaggeration. In our simulations, we evaluate gender diversity using the "male" vs "female" partition and skintone diversity using the Fitzpatrick scale. Pre-defined protected attribute partitions, however, can be prob-

lematic; e.g., commercial AI tools' inability in handling non-binary gender [269].

Considering that Our algorithm is based on choosing control sets that can differentiate across protected attribute types, there is a possibility that the automatically constructed control sets can be stereotypically biased. For example, a control set with a high $\gamma^{(T)}$ value for gender may just include images of men and women, and exclude images of transgender individuals. While *non-redundancy* aims to ensure that the control set is diverse, it does not guarantee that the control set will be perfectly representative. Given this possibility, we strongly encourage the additional hand-curation of automatically-constructed control sets. Further, any agency using control sets should make them public and elicit community feedback to avoid representational biases. Recent work on designs for such cooperative frameworks can be employed for this purpose [221, 109].

Choice of α. For *DivScore-Adaptive*, α is the parameter that controls the redundancy of the control set. It primarily depends on the domain in consideration and we use fixed α for collections from the same domain. However, the mechanism to choose the best α for a given domain is unclear and can be further explored.

Improving theoretical bounds. While the theoretical bounds provide intuition about the dependence of error on the size of the control set and γ, the constants in the bounds can be further improved. E.g., in the case of the PPB dataset with gender protected attribute and the empirical setup in Section 3.4.1, Theorem 3.3.3 suggests that error $|\delta| \leq 5$; however, we observe that the error is much smaller (≤ 0.5) in practice. Improved and tighter analysis can help reduce the difference between the theoretical and empirical performance.

Assessing qualitative disparities. Our approach is more cost-effective than crowd annotation. However, crowd-annotation can help answer questions about the col-

lection beyond disparity quantification. For example, Kay et al. [166] use crowd-annotation to provide evidence of sexualized depictions of women in Google Image results for certain occupations such as construction worker. As part of future work, one can explore extensions of our approach or control sets that can assess such qualitative disparities as well.

The use of control sets (or small sets of representative examples) allows us to audit for biases in the absence of protected attributes. But representative examples here have a larger role: they are a general context-specific signal of a group membership. In the case of images, representation from the perspective of the user is the appropriate depiction of images containing people with diverse perceived attributes. For instance, race obviously cannot be inferred using images but skintone is a signal that is often used by people to determine race representation in image sets. Representative examples make it easier to incorporate these perceived signals of protected attributes. Secondly, control sets here can be defined by each user themselves, allowing them to define their notion of diversity through examples. On this point, considering that the above process allows us to audit efficiently, it should also be possible to diversify datasets so that they *appear similar* to the control set defined by any user. In particular, Chapter 4 and Chapter 5 demonstrate how control sets can be used to diversify image and text summaries so that they represent all groups in a similar manner as any given control set.

Chapter 4

Implicit Diversity in Image Summarization

Services such as Google Image Search perform the task of image summarization; namely, responding to a query with an appropriate set of images. However, as mentioned in Chapter 3, for queries related to people, such algorithms are often biased with respect to protected attributes of the data, such as the presented gender [166, 278] or skin tone [39]. In essence, summarization algorithms often overrepresent the majority demographics for a given query. Kay et al. [166] show that such errors can reinforce the gender stereotypes associated with common queries, underlining the need to correct such biases in image summarization results. Furthermore, the use of demographically skewed results can be propagated and reinforced by other tools; e.g., state-of-the-art image generation algorithms such as Generative Adversarial Networks (GANs), when trained on publicly available images of engineers, mostly generate images of white men wearing a hard hat [7].

This chapter is based on joint work with L. Elisa Celis and was published in the proceedings of ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW) 2020 [50]. I would like to thank the anonymous area chairs and reviewers of CSCW'20 for their thorough and helpful feedback.

Clearly, there is a necessity for developing image summarization algorithms that do not propagate or exacerbate societal biases and that generate summaries that are relevant to the given query yet are also visibly diverse.

Most existing approaches for fair and diverse summarization assume that the images of people include labels denoting the relevant protected attributes of individuals in the images. These labels are explicitly used to either change the dataset or adjust the training of the summarization algorithm. However, such labels are often unknown, as in the case of images in Google Search results. Further, using machine learning techniques to infer these labels may often not be possible within acceptable accuracy ranges and may not be desirable due to the additional biases this process could incur.

This chapter presents a novel approach that takes as input a visibly diverse control set of images of people and uses this set as part of a procedure to select a summary of images of people in response to a query. Extending the use of control sets from Chapter 3, the goal is to have a resulting summary that is more visibly diverse in a manner that emulates the diversity depicted in the control set. Our algorithms accomplish this by evaluating the similarity of the images selected by a black-box algorithm with the images in the control set, and incorporating this "diversity score" into the final selection process. Importantly, this approach does not require images to be labeled at any point; effectively, it gives a way to implicitly diversify the set of images selected.

Summary of contributions. In 2013-14, Kay et al. [166] collected Google's top 400 image results for each of the 96 occupations, and had 10% of the images labeled by crowd workers according to presented gender. They used this dataset to infer the gender bias in the Google search results of occupations described above. In the years since then, Google has continually updated its image analysis algorithms

[3]. Hence, the first question we address is: *does bias remain an issue in Google image search results?*

Towards this, we consider the same 96 occupations and collect the top 100 Google search results for each one in December 2019.[1] We have these images labeled by crowd workers using Amazon Mechanical Turk (AMT) with respect to gender (coded as male, female, or other) and skintone (coded according to the Fitzpatrick skin-tone scale). This results in 60% of images containing gender labels and 63% of images containing skin-tone labels. While some improvements have been made with respect to gender (the % of images of women in Google 2014 results is 37% and in Google 2019 results it is 45%), we find that the fraction of gender anti-stereotypical images is still quite low[2] (30% in Google 2019 results and 22% in Google 2014 results).

For skintone, 52% of the images have a *fair* skin-tone label (corresponding to Type 1-3 on the Fitzpatrick scale) and 10% of the images have a *dark* skin-tone label (corresponding to Type 4-6 on the Fitzpatrick scale). Once again, the fraction of images of dark-skinned people in Google results is quite low. Overall 57% of the dataset has both a gender and skin-tone label; however, only 7% of these are images of dark-skinned men and 3% are images of dark-skinned women. A final statistic that captures the lack of diversity in Google results is that 35 out of 96 occupations do not have any images of dark-skinned gender anti-stereotypical people in the top 100 results. This assessment of Google images with respect to skintone was not possible for the original dataset of images from 2014, as no skintone labels were present.

Given the extent and importance of this problem, the next question we address

[1] Dataset available at http://bit.ly/2QVfM0K

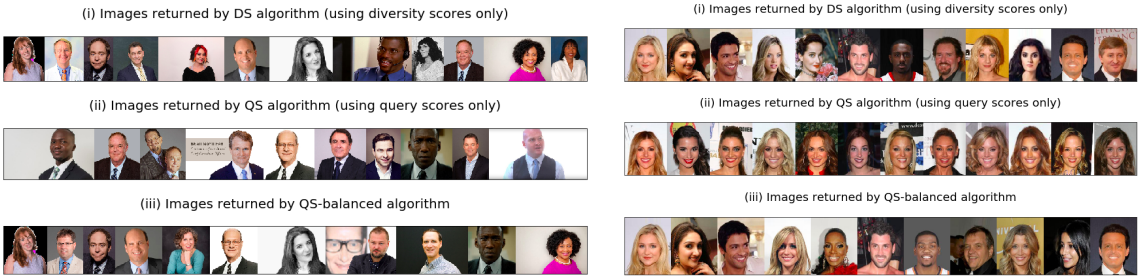
[2] Anti-stereotypical images refer to a set of images that do not correspond to the stereotype associated with the query. For example, gender anti-stereotypical images for a male-dominated occupation (determined using ground truth) would correspond to the set of images of women in the summary generated for that occupation.

is: *are there simple and efficient methods that correct for visible diversity across protected attributes in image search?* When considering this question, we first note that, in general, images that contain people would not have their protected attributes explicitly labeled. Datasets are at scales where collecting explicit labels is infeasible, and while it may be possible to *learn* these attributes in a pre-processing step, as we also observe this can lead to additional errors and biases [269]. Hence, we add a constraint to our main question: *are there simple and efficient methods that correct for visible diversity across protected attributes in image search results that do not require or infer attribute labels?* To the best of our knowledge, no methods with such a requirement exist for image summarization. [3]

To address this question, we design two algorithms: *MMR-balanced*, a modification of the well-known MMR algorithm [46], and *QS-balanced*, a simpler and more efficient algorithm inspired by the former. In both cases, the method takes a black-box image summarization algorithm and the dataset it works with, and overlays it with a post-processing step that attempts to diversify the results of the black-box algorithm. To do so, our method takes as input a very small control set of visibly diverse images. The control set is query-independent and should be carefully constructed to capture the kind of visible diversity desired in the output. Similar to Chapter 3, control sets here encode the user's notion of diversity. [4] On a high level, the process of debiasing summaries using control sets is as follows (see also Figure 4.2): each image is given a query similarity score using the black-box algorithm, which corresponds to how well it represents the desired query. The

[3] The goal of search algorithms is usually to return a ranking of images given an input query. While our approach can be extended to the case of ranking as well, in this chapter, we will primarily focus on the task of fair retrieval, i.e., returning a fair *summary* of images corresponding to an input query and ensuring that the top results are unbiased. The reason for this simplification is to better analyze, highlight and mitigate the bias in the most visible results of image search, often characterized by images on the first or second page of the search results. However, as discussed in Remark 4.2.1, our algorithms can be used to rank images in a diverse manner as well.

[4] The size of the control set can vary by application, but we show the efficacy of our method with small sets of size 8-25.



(a) Occupations dataset- Query: CEO

(b) CelebA dataset- Query: Smiling

Figure 4.1: (a) Top images returned by *QS-balanced* for the query "CEO" on Occupations dataset and (b) top images returned by *QS-balanced* for the query "smiling" on CelebA dataset. The first row shows images returned by the algorithm using the diversity control matrix, the second row shows the images with the most similarity to the query, and the third row shows images with best-combined scores, i.e., images with smallest $\{\text{DS}_q(x, x')\}_{x \in S}$ scores for each $x' \in T$.

candidate images are also given a similarity score with respect to each image in the control set using a given similarity scoring tool. After adding the query similarity score to the diversity control scores, we rank the images by the combined score for each image in the control set and output the ones with the best scores. As required, this results in a method that implicitly diversifies the image sets without having to infer or obtain protected attribute labels.

We evaluate the effectiveness of this approach on the new Occupations dataset we collect and the CelebA dataset. The CelebA dataset contains more than 200,000 images of celebrities labeled with information about the facial attributes of the person in the image. For the Occupations dataset, the queries are the occupations while for the CelebA dataset, the queries are the facial attributes.

We compare the performance of our approaches on these datasets with other state-of-the-art algorithms and relevant baselines. This includes summarization algorithms that reduce redundancy in the summary [46], diversify across the feature space [177], or use gender classification tools to compute explicit labels as a preprocessing step. For the Occupations dataset, *QS-balanced* and *MMR-balanced* return more gender-balanced results than Google image search results (Section 4.4.3)

and baselines. Specifically, the percent of gender anti-stereotypical images in the output of *QS-balanced* and *MMR-balanced* is around 45% on average across occupations, while for Google Image search, this number is approximately 30%. The baseline algorithms also have a relatively lower percent of gender anti-stereotypical images in their output (35%-39%), confirming observations made in prior work which state that diversifying across feature space or using pre-trained gender classification tools do not necessarily result in diversity with respect to protected attributes [51, 269]. Similarly, on the CelebA dataset, our algorithms return much more gender-balanced results, compared to the results using just query similarity or other algorithms. In this case, the average fraction of gender anti-stereotypical images in the output of *QS-balanced* is 0.23, while using just query similarity, this number is 0.08. For example, for gender-neutral facial attributes, such as 'smiling', the 50 images obtained using top query scores are images of women while *QS-balanced* returns an image set with 32% men and no loss in accuracy. On the Occupations dataset, we also show that *QS-balanced* and *MMR-balanced* increase the diversity across skintone as well as diversity across the intersection of skintone and gender.[5] The average fraction of images of dark-skinned people in the output *QS-balanced* is 0.17 while for Google results, the average fraction is 0.16. However, the standard deviation is higher for Google results (0.09 vs 0.05), implying that the results are relatively more unbalanced for Google. In terms of intersectional diversity, Table 4.1 shows that the results from *QS-balanced* algorithm are gender-balanced across skintone, unlike Google results. The average fraction of images of dark-skinned gender anti-stereotypical images in the output of *QS-balanced* is 0.08 while for Google, this number is 0.05. The increase in diversity with respect to skintone is limited, perhaps due to the lack of skin-tone diversity in the dataset itself. We show that we can improve on these numbers by more aggressively weight-

[5]The CelebA dataset does not contain race or skin tone labels, hence we cannot evaluate its performance with respect to these attributes.

Table 4.1: Comparison of intersectional diversity of top 50 *QS-balanced* images and Google images. The number represents the average fraction of images satisfying the corresponding attribute, with standard deviation in the brackets.Google images seem to have a larger fraction of stereotypical images, with respect to both gender and skintone. In comparison, *QS-balanced* returns images that are relatively more balanced; for both skintones, the fraction of men and women in the output is almost balanced. Intersectional diversity comparison with other baselines is presented in Table A.1.

	Our Algorithm	
	% gender stereotypical	% gender anti-stereotypical
Fair skin	0.46 (0.14)	0.37 (0.14)
Dark skin	0.09 (0.05)	0.08 (0.05)
	Google Images	
	% gender stereotypical	% gender anti-stereotypical
Fair skin	0.60 (0.20)	0.24 (0.21)
Dark skin	0.11 (0.08)	0.05 (0.07)

ing the diversity score (computed with respect to the control set), this comes at an increased cost to accuracy.

Importantly, our focus in this chapter is on visible diversity with respect to perceived gender and skin color. We make this choice as true labels are often not only unknown but also irrelevant – e.g., a set of images of male-presenting CEOs is not sufficiently diverse to combat the problems mentioned above, regardless of the true gender identity of the people captured in the images. As discussed in Chapter 2, how we define appropriate representation and diversity can be highly context-dependent; it can either be used to mean *fidelity* with ground truth or can denote that there are sufficient number of samples corresponding to each relevant demographic group. In this Chapter, our analysis focuses on both of these aspects of representation. We compare the gender and skintone diversity in Google Search results and summaries generated by our algorithms for various occupations to the actual demographic distribution in these occupations in the US using sur-

vey data from Bureau of Labor and Statistics, to measure deviation of these summaries from reality. Simultaneously, we also quantify the extent to which stereotypes associated with various occupations are propagated or exaggerated in the automatically-generated summaries, giving us an idea of the under-representation of historically-marginalized groups in image summaries.

The following sections are organized as follows: after briefly reviewing related work in the field of diverse image summarization, we start with a description of the setting of summarization, followed by the details of our suggested algorithms in Section 4.2. We next present the Occupations dataset and assess the gender and skin-tone diversity of the dataset in detail in Section 4.3. Following this, we state the results of the empirical analysis of our algorithm on the Occupations and CelebA dataset (Section 4.4). Finally, we discuss the implications and inferences from our results and address the limitations of our methods and ways to improve them in future work (Section 4.5).

4.1 Related Work

To assess the importance of addressing bias in summarization results, we first look at prior work on the social impact of stereotypes in image datasets and related work in the field of fair summarization.

Bias in existing image datasets and models. The effect of negative stereotypes and the resulting biases have been carefully explored in television media in the form of *cultivation theory* [275, 117], particularly with respect to the portrayal of women, racial and ethnic minorities. Online media has only recently been subjected to similar scrutiny and multiple studies have highlighted the presence of such biases in existing summarization tools and benchmark image datasets.

As discussed before, the study by Kay et al. [166] explored the effects of bias

in Google image search results of occupations on the perception of people of that occupation. Follow-up studies by Pew Research Center [5] and Singh et al. [278] also found evidence of similar gender bias in Google image search results; [5] further observed that, for many occupations, images of women tend to appear lower than the images of men in search results. Biased representation of minorities has also been observed in other computer vision applications. Buolamwini and Gebru [39] found that popular facial analysis tools from IBM, Microsoft, and Face++ have a significantly larger error rate for dark-skinned women than other groups. This study led to a subsequent improvement in the accuracy of these tools with respect to images of minorities [4] and it highlights the importance of constant audit of existing models, as well as, the need for alternative strategies to develop unbiased models since even improvements to existing facial analysis tools do not achieve desired diversity in their results. A case in point is the study by Scheuerman, Paul, and Brubaker [269] which showed that commercial facial analysis tools do not perform well for transgender individuals and are unable to infer non-binary gender.

Even existing datasets, collected from real-world settings, can encode unwarranted biases that can occur from the data collection process. Van Miltenburg [296] provided evidence of stereotype bias in a popular dataset of Flickr images annotated with crowdsourced descriptions. The study by Zhao et al. [320] found that datasets used for visual recognition tasks have a significant gender bias.

Downstream propagation of biases. As mentioned earlier, inaccurate representation of demographic groups can lead to biases against these groups, either in the form of incorrect perceptions about the group [166, 68, 138] or in the form of bias in the decision-making process based on the inaccurate representations. [247, 74, 255, 23, 163]. If a machine learning model is trained using an imbalanced or misrepresentative dataset, the biases in the dataset can edge into the output of

the model as well. For example, Datta et al. [79] showed that men are more likely to be shown Google ads for high-paying jobs than women, a result of training the targeting model on gender-biased data. Similarly, Caliskan et al. [44] found that word associations learned from existing texts encode historical biases, such as gender stereotypes for occupations. Image generation algorithms, such as GANs [164], when trained on Google Images of people from certain common occupations, mostly generate stereotypical images [7]. With any additional intervention, unconstrained models, including summarization algorithms, are bound to reflect the biases of the dataset they operate upon. Hence, to prevent the propagation of bias due to imbalanced image summaries, it is important to develop summarization algorithms that ensure that the generated summaries are unbiased even when using biased datasets.

Algorithms for image summarization. The rising popularity of social networks and image-hosting websites has led to a growing interest in the task of image summarization. The primary goal of any image summarization algorithm is to appropriately condense a given set of images into a small representative set. This task can be divided into two parts: (a) scoring images based on their importance, and (b) ensuring that the summary represents all the relevant images.

Traditional image summarization algorithms to score images on their importance have focused on using visual features, such as color or texture, to compare and rank images [132, 310]. Recently, even pretrained neural networks have been used for image feature extraction [274], which is then used to score images based on their *centrality* in the dataset. In the case of query-based summarization, determining the importance of an image includes determining whether the image is relevant to the query. To find query-relevant images, search services like Google use metadata from the parent websites of images to associate keywords with them,

thus simplifying the task significantly [311]. However, for the datasets we analyze, metadata or keywords for images are not available; correspondingly we need to use retrieval algorithms that use image features only. If the queries come from a pre-determined set, then supervised approaches for image classification can also be used for summarization [263, 256, 317, 100]. For example, if the queries correspond to facial features, then scores from state-of-the-art convolutional neural networks pre-trained on large image datasets with annotated facial features [199] can be employed for retrieving relevant images. We will show the efficacy of such an approach in Section 4.4 for the CelebA dataset. In the absence of pre-trained classification models and metadata information, one has to adopt unsupervised approaches to determine the query relevance of images. Given a query image, an unsupervised approach suggested by [303] uses pre-trained models [85] to find images similar to the query image; they show that this unsupervised approach is comparable to state-of-the-art algorithms for the task of *pattern spotting*. We will use this approach for query-based summarization for the Occupations dataset.

Secondly, to ensure that the summary is representative of all relevant images, most prior works have used the idea of *non-redundancy* [46, 251, 266, 66, 193]. Once the images have been scored on their relevance, algorithms such as MMR [46], greedily select images that are not very similar to the images already present in the summary. Other efficient methods to ensure non-redundancy in the summary include the use of determinantal point processes [177] and submodular maximization models [294]. These models have also been used explicitly for the task of efficiently summarizing images of people [279]. However, reducing redundancy in the output set does not always correspond to diversity with respect to the desired features, such as gender, race, etc., as demonstrated by Celis et al. [51]. Our evaluations using redundancy-reducing algorithms also lead to this conclusion. We empirically compare our algorithm to such non-redundancy-based approaches in

Section 4.4 and discuss them further in Section 4.5.

Prior work on unbiased image summarization. Current approaches to debias summarization algorithms often assume the existence of protected attribute labels for data points. Lin et al. [193] suggest a scoring function over subsets of elements that rewards subsets that have images from different partitions. For example, Celis et al. [52] formulate the summarization problem as sampling from a Determinantal Point Process and use partition constraints on the support to ensure fairness. However, setting up the partition constraints or evaluating scores requires the knowledge of the partitions and correspondingly the protected attributes for all data points. Similarly, fair classification algorithms, such as [54, 70, 98, 136, 161, 312, 316] use the gender labels during the training process. Even for language-based image recognition tasks, [320] suggest constraints-based modifications of existing models to ensure fairness of these models, but the constraints are based on the knowledge of the gender labels. Unlike these approaches, the methods proposed in this chapter aim to ensure diversity in settings where protected attribute labels are not available.

4.2 Model and Algorithms

In this section, we describe our approach to ensuring that the image summarization process returns visibly diverse images. Given a query from the user, we start with the goal of choosing images that correspond to the query and then incorporate an additional novel diversity check (using a control set provided by the user) into the model. Let S denote the large corpus of images.

Query Score. Suppose we have a black-box algorithm A that takes any query q and the dataset S as input and for each image, returns a query similarity score - the

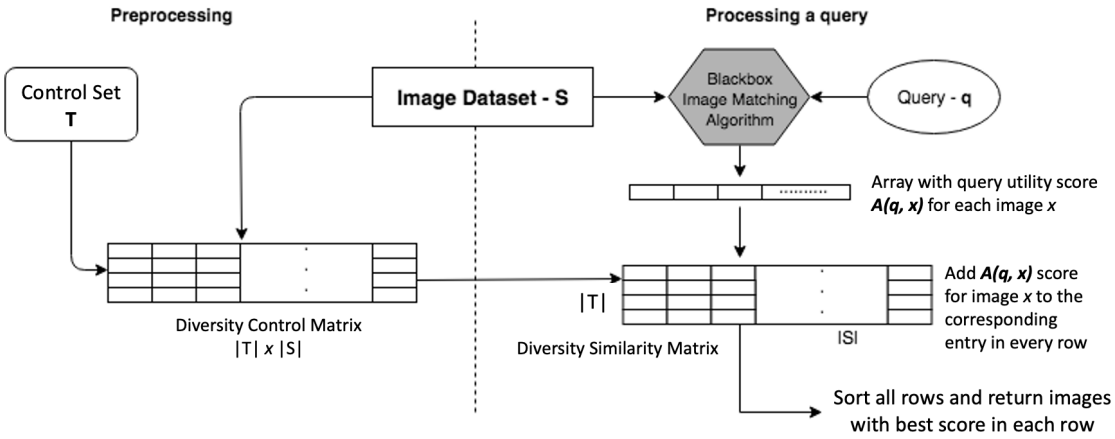


Figure 4.2: A simple post-processing approach for ensuring diversity in image search. A small "control set" of images is taken as input, and (relevant) images are assigned a similarity score with each image in the control set to create a diversity control matrix. These scores are combined with the query scores provided in a black-box manner using an existing image search approach. A summarization algorithm then selects the final images using this combined score. See Algorithm 4 for details.

score represents how well the image corresponds to the query q. The smaller the score $A(q, x)$, for a query q and image x, the better the image corresponds to the query. Since our framework is meant to extend an existing image retrieval model, we can assume that such a score can be efficiently computed for each query and image pair.

Image Similarity Score. Suppose that we also have a generic image similarity function $\text{sim}(\cdot, \cdot)$, which takes as input a pair of images, x_1, x_2, and calculates a score of similarity of the two images, $\text{sim}(x_1, x_2)$. For the sake of consistency, here again, we will assume that the smaller the score, the more similar are the images.

While the framework we propose is independent of the query-matching algorithm or the image similarity function, we will present a concrete example of such algorithms and functions in a later section. We first see how we can use this score to rank our dataset.

Algorithm 3 MMR-balanced

Input: Dataset S, query q, query matching algorithm A, similarity function sim, control set T, parameter $\alpha, \beta \in [0, 1]$, number of elements to be returned $M < |S|$
Output: Summary R

$R \leftarrow \emptyset$
while $|R| < M$ **do**
  for all $x \in S \setminus T_R$ **do**
    redundancy-score $\leftarrow \min_{x' \in T_R} \text{sim}(x, x')$
    diversity-score $\leftarrow \min_{x_c \in T} \text{sim}(x, x_c)$
    score$(x) \leftarrow (1 - \alpha - \beta) \cdot A(q, x) - \beta \cdot$ redundancy-score $+ \alpha \cdot$ diversity-score
  $R \leftarrow R \cup \arg\min_x \text{score}(x)$
return R

Diversity using a control set. A ranking/summary with respect to the scores returned by A is unlikely to be visibly diverse without further intervention in most cases, as shown by prior studies [166]. To ensure visible diversity in the results, we use a control set T and a clustering approach. Like Chapter 3, the control set T is a *small* set of visibly diverse images and will be used to enforce the diversity in the output; for example, if the summary is required to be gender-diverse, then the control set will have an equal number of images of men and women.

For each control image $x_c \in T$, using $\text{sim}(\cdot, \cdot)$ as the distance metric, we can learn the cluster of images around x, by sorting $\{\text{sim}(x, x_c)\}_{x \in S}$ for each $x_c \in T$. In other words, we can associate each image $x \in S$ to an image in the control set to which x is most similar.

Using control sets with existing redundancy-reducing algorithms. To ensure we take into account both the query score from the blackbox A and the diversity with respect to the control set T, we have to combine the scores $A(q, \cdot)$ and $\text{sim}(x, \cdot)$. As mentioned earlier, a popular approach to combining query similarity and diversity is to diversify across the entire feature space, i.e., reduce the *redun-*

dancy of the chosen summary. Using *maximum marginal relevance* score is one of the many simple and efficient greedy selection procedures for this task [46]. The maximum marginal relevance (MMR) score of an image is a combination of the query similarity score of that image and its dis-similarity to the already chosen images; at every step, the image that optimizes this score is added to the set. However, reducing redundancy does not necessarily lead to diversification across the desired attributes, such as gender [51]. An obvious question in this respect is whether the control set score can be incorporated with a non-redundancy approach to achieve diversity across gender, race, etc.

To that end, we present the *MMR-balanced* algorithm. Starting with an empty set R, the algorithm adds one image to the subset R in each iteration. The chosen image x is the one that minimizes the score

$$(1-\alpha-\beta)\cdot A(q,x) - \beta \cdot \min_{x' \in R} \text{sim}(x,x') + \alpha \cdot \min_{x_c \in T} \text{sim}(x,x_c),$$

where $\alpha, \beta \in [0,1]$. The first part of the above expression captures query relevance while the second part penalizes an image according to similarity to existing images in the summary R. These two terms together constitute the *maximum marginal relevance* score [46]. The third term in the above expression now acts as a deterrent to choosing multiple images corresponding to the same control set image x_c (unless there is an almost equal number of images corresponding to each x_c in R). The complete algorithm is formally presented in Algorithm 3. We will set $\alpha = \beta = 0.33$ for *MMR-balanced* in the following sections and empirical analysis. We also analyze this expression theoretically in Appendix A.2.3.

A drawback of *MMR-balanced* is the time complexity. In particular, checking redundancy with existing images at every step is cumbersome and often necessary, if the dataset is diverse enough. Furthermore, dropping the redundancy check

Algorithm 4 QS-Balanced: Post-processing algorithm for fair summarization

Input: Dataset S, query q, blackbox algorithm A, similarity function $\text{sim}(\cdot, \cdot)$, control set T, parameter α, and summary size M
Output: Summary R

for all $x, x_c \in S \times T$ **do**
    $\text{DS}_q(x, x_c) \leftarrow (1 - \alpha) \cdot A(q, x) + \alpha \cdot \text{sim}(x, x_c)$
$R \leftarrow \emptyset$
while $|R| < M$ **do**
    $r, \text{score} \leftarrow \emptyset$
    for all $x_c \in T$ **do** ▷ Find elements clustered around each x_c
        $x' \leftarrow \arg\max_{x \in S} \text{DS}_q(x, x_c)$
        if $x' \notin r$ **then** ▷ Checking duplicates
            $r \leftarrow r \cup \{x'\}$
            $\text{score}(x') \leftarrow \text{DS}_q(x', x_c)$ ▷ Scores used for tie-breaks
            $\text{DS}_q(x', x_c) \leftarrow -\infty$
    if $|R \cup r| \leq M$ **then** ▷ If all of r can be added
        $R \leftarrow R \cup r$
    else ▷ Tie-break when $|R \cup r|$ has more than M elements
        $m' \leftarrow M - |r|$
        $r' \leftarrow m'$ elements from r with highest $\text{score}(x')$
        $R \leftarrow R \cup r'$
return R

should not affect the diversity with respect to protected attributes, since we have the diversity control term for that purpose. This leads us to a more efficient algorithm.

QS-balanced. Given a tradeoff parameter $\alpha \in [0, 1]$ and a query q, for each $x_c \in T$ let $\text{DS}_q(x, x_c) : S \times T \to \mathbb{R}$ denote the following score function:

$$\text{DS}_q(x, x_c) \leftarrow (1 - \alpha) \cdot A(q, x) + \alpha \cdot \text{sim}(x, x_c).$$

The score $\text{DS}_q(x, x_c)$ corresponds to a combination of similarity with x_c and similarity with a query q. Finally, for each $x_c \in T_F$, we sort the set $\left\{\text{DS}_q(x, x_c)\right\}_{x \in S}$ and return an equal number of images with the lowest scores from each set, check-

ing for duplicates at every step. The ties are broken by choosing the image with the better query score. This gives us our final set of visibly diverse images. Algorithm 4 formally summarizes this approach. For $\alpha = 0.5$ and given a control set, we will call this algorithm *QS-balanced*. We will also refer to the algorithm using only diversity scores, i.e., $\alpha = 1$, as *DS* and the algorithm using only query scores, i.e., $\alpha = 0$, as *QS* in the following sections.

Time complexity of the QS-balanced. Without making any assumption on the blackbox algorithm A, we can upper bound the additional time to ensure diversity using the control set. The additional overhead in time complexity is $\mathcal{O}(|T| \cdot \log(|S|) \cdot \mathcal{T})$, where $\mathcal{T}$ is the time taken to compute the similarity score for any given pair of elements. This factor is due to the time taken to construct and sort the rows of the diversity-similarity matrix. The time complexity also depends linearly on the size of the control set and hence the size of the control set should be much smaller than the size of the dataset. Note that *MMR-balanced* is $O(M)$ times slower than *QS-balanced*, where M is the size of the summary.

Model Properties. An important property that many diverse summarization algorithms (including *MMR*) share is the *diminishing returns property* [46, 193, 294]. To state briefly, a function, defined over the subsets of a domain, satisfies the *diminishing returns property* if the change in function value on adding an element to a smaller set is relatively larger. Such set functions are also called *submodular functions*. Due to diminishing returns property, simple greedy algorithms can be used to approximately and efficiently optimize these functions, making them ideal for summarization over large datasets.

We can directly show that score computed at each step of *MMR-balanced* satisfies the diminishing returns property (simple extension of proof for *MMR*). Even *QS-balanced*, if represented as an iterative process, can be shown to satisfy this

property, implying that these algorithms share the mathematical features of common diverse summarization algorithms and that fast and greedy approaches do lead to approximately good solutions. We formalize these statements and provide mathematical proofs of the submodularity of these functions in Appendix A.2.3.

Remark 4.2.1 (Ranking)**.** *Search algorithms usually return a ranking of images in the dataset and ranking models also suffer from the same kind of biases studied in the case of summarization [5, 53]. While ranking a set of images can be considered an extension of the summarization problem, we primarily focus on summarization to highlight and mitigate bias in the most visible results of image search. However, given the similarity between these problems, an obvious question is whether our approach can be used to provide a fair ranking of the images. Indeed, both QS-balanced and MMR-balanced can be used to rank images as well. Both algorithms inherently compute a score for each image which captures both the query similarity and diversity with respect to the control set (see Section A.2.3 for more details). While the QS-balanced is for diverse image summarization, with slight modification the algorithm can also be used to rank the images in the dataset according to the score $DS_q(x, x_c)$. We can construct a $|S| \times |T|$ sized matrix (as shown in Figure 4.2) with the entry corresponding to (x, x_c) storing the score $DS_q(x, x_c)$. Next, we first sort each row of this matrix according to the stored score and then sort each column. Finally, we can assign a ranking, starting with the image corresponding to the first entry of the matrix and moving along the first column. Once the first column has been ranked, we move to the second column and so on, checking for duplicates at each step.*

4.3 Datasets

4.3.1 Occupations Dataset

We compile and analyze a new dataset of images for different occupations. The dataset is composed of the top 100 Google Image Search results [6] for 96 different occupations. This dataset is an updated version of the one compiled by Kay, Matuszek and Munson [166], which contained Google image results from 2013 [7].

Since occupations are often associated with gender or race stereotypes, empirical analysis with respect to these search terms will help better evaluate the imbalance in existing search and summarization algorithms. To compare the composition of the dataset with the ground truth of the fraction of minorities working in the occupation, we use the census data of the fraction of women and Black people working in each occupation from the Bureau of Labor and Statistics [2]. The census data shows that Black people are the racial minority in each of the considered occupations (relative to White people). On the other hand, 52 out of 96 occupations have a larger fraction of men employed and the rest have a larger fraction of women employed. In our analysis, we will often compute the fraction of gender anti-stereotypical images for different occupations, i.e., if an occupation is male-dominated, we take into account the fraction of women and if an occupation is female-dominated, we take into account the fraction of men in the output set.

We use Amazon Mechanical Turk to label the gender and Fitzpatrick skintone of the primary person in the images. To obtain labels, we designed a survey asking participants to label the gender and skintone of the primary person in the images. Each survey had around 50 images and the surveys were limited to participants in the US. Since some of the images had multiple primary persons or people whose features were hidden or cartoon images, "Not applicable" and "Cannot

[6] The images were collected in December 2019.

[7] `https://github.com/mjskay/gender-in-image-search`

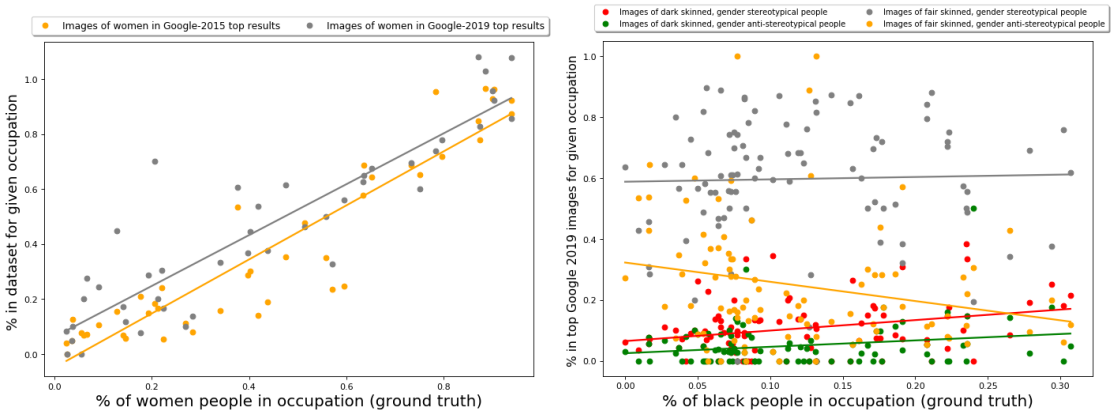
determine" were also provided for each question. For each image, we collect 3 responses and assign the majority label to the image.

We use standard inter-rater reliability measurements to quantify the extent of consensus amongst different participants of the survey. Overall there were around 620 survey participants and each participant only labels a small subset of images (50). We compute the Cohen's κ-coefficient [67] for all pairs of participants with more than 5 common images in their surveys. [8] The resulting mean κ-coefficient across the pairs is 0.58 (median is 0.62). Based on existing heuristic guidelines and interpretations of these coefficients [179], these results imply that, on average, there is a *moderate* level of agreement between survey participants.

An analysis of this dataset revealed similar diversity results as the analysis by Kay et al. [166] of Google images from 2013. However, while their analysis was limited to gender, we are also able to assess the skin-tone diversity of the results. Furthermore, unlike Kay et al., who mainly report the fraction of images of women in top results, we focus on measuring the fraction of gender anti-stereotypical images in top images. This is because our primary goal is to provide balanced summaries and present anti-stereotypical images to effectively counter gender stereotypes [109]. Measuring the fraction of anti-stereotypical images better quantifies the stereotype exaggeration in current results, compared to the fraction of images of women.

Gender labels. Overall, approximately 61% of these images have a primary person whose gender is labeled as either Male or Female. 35% of the images are labelled *Male*, 26% are labelled *Female* and the rest are labelled "Not applicable" and "Cannot determine". The variation of the fraction of images of women in the results is presented in Figure 4.3a. The figure shows that Google images do follow

[8] Similar techniques to evaluate interrater agreement in the setting of multiple participants rating a subset of elements has been considered in other prior work as well [189, 220].



(a) Fraction of images of women in Google results

(b) Fraction of images of dark-skinned people and dark-skinned women in Google results

Figure 4.3: The plots show the fraction of images of women, dark-skinned people, and their intersection in the top 100 results of Google Image Search. (a) For gender, we also provide the comparison with Google results from 2013 [166]. While the fraction of women in the top Google results seems to have increased, the fraction of gender-stereotypical images is still high (0.7 on average). (b) Majority of the top Google images for every occupation correspond to gender-stereotypical fair-skinned people, independent of the ground truth of the percentage of Black people in the occupation. For the rest of the minority groups, the fraction is partially dependent on the ground truth.

the gender stereotype associated with occupations. This was one of the main inferences of the case study by Kay et al. [166] for Google 2013 search results. While the overall fraction of women in the top 100 results seems to have increased from 2013 to 2019 (37% in 2013 to 45% in 2019), the fraction of gender anti-stereotypical images is still quite low (21% in 2013 and 30% in 2019).

Skin-tone labels. For skintone, the options provided for labeling were the categories of the Fitzpatrick skin-tone scale (Type 1-6). While there are more options, in this case, choosing between consecutive options is relatively difficult. Around 15% of the images are assigned a Type-1 skin-tone label, 14% Type-2, 5% Type-3, 2% Type-4, 2% Type-5, 2% Type-6; the rest are either "Not applicable", "Cannot determine" or have conflicting skin-tone label responses.

However, our primary skin-tone evaluation is with respect to the fraction of

Table 4.2: Occupations dataset - Comparison of top 50 images from *QS-balanced* and *MMR-balanced* algorithm with top 50 images from other baselines. The number represents the average, with the standard deviation in brackets. The accuracy is quantified using a measure of similarity to the query. *QS-balanced* returns an output set that has a larger fraction of images that do not correspond to the gender stereotype of the occupation. However, it suffers a loss in accuracy for this diversification. Note that accuracy, in this case, is measured using query similarity. Other non-redundancy-based algorithms also perform better than Google results in terms of gender diversity in the results, but not better than *QS-balanced* or *MMR-balanced*, showing that using the control set targets the desired attributes better.

Algorithm		**Diversity metrics**		**Accuracy metric**
		% gender anti-stereotypical	% dark skinned	avg. accuracy
Our algorithms	QS-balanced	0.45 (0.17)	0.17 (0.05)	0.38 (0.06)
	MMR-balanced	0.45 (0.20)	0.15 (0.06)	0.39 (0.06)
Baselines	QS	0.35 (0.20)	0.13 (0.06)	0.47 (0.11)
	DS	0.48 (0.20)	0.15 (0.00)	0.30 (0.06)
	Google	0.30 (0.22)	0.16 (0.09)	0.48 (0.07)
	MMR	0.35 (0.21)	0.09 (0.05)	0.48 (0.11)
	DET	0.39 (0.15)	0.15 (0.05)	0.43 (0.08)
	AUTOLABEL	0.36 (0.17)	0.14 (0.05)	0.47 (0.11)
	AUTOLABEL-RWD	0.35 (0.21)	0.13 (0.06)	0.47 (0.11)

images of dark-skinned people. Hence we can aggregate the skintones into a binary feature: *fair* skintone (Type 1, Type 2, Type 3) and *dark* skintone (Type 4, Type 5, Type 6). After this aggregation, 52% of the images have the fair skin-tone label and 10% of the images have the dark skin-tone label. For the rest of the chapter, we will treat the skintone as a binary feature, unless explicitly mentioned.

Intersection of gender and skintone. 57% of the images have both a gender and skin-tone (binary) label. Amongst these, 27% of the images are of fair-skinned men, 21% are of fair-skinned women, 6% are of dark-skinned men and 3% are of dark-skinned women. Once again, the fraction of images of dark-skinned men and women is relatively much smaller than the fraction of fair-skinned men and women, as seen from Figure 4.3b. Furthermore, if we associate each occupation

with its gender stereotype (for example, "Male" if the fraction of men in the occupation is larger than the fraction and women, and "Female" otherwise), then 35 out of 96 occupations do not have any images of dark-skinned gender anti-stereotypical people in the top 100 results.

Figure 4.3b also provides us with an insight into the variation of the fraction of images of different groups (formed by the intersection of gender and skintone) with respect to the ground truth of the fraction of Black people in occupations. For almost all occupations, a large portion of the top 100 images is of gender-stereotypical fair-skinned people, further showing that current Google results for occupations do correspond to the stereotypes. Interestingly, the fraction of images of gender-stereotypical fair-skinned people do not seem to be dependent on the ground truth. While this partition takes up a significant portion of the top 100 images, the fraction of images from the other three minority partitions seems to be partially dependent on the ground truth.

This lack of gender diversity in Google results from 2013 has also been explored in detail in the paper by Kay et al. [166]; our updated dataset shows that the current Google results still suffer from some of the gender diversity problems discussed in Kay et al. [166]. Furthermore, our analysis also shows that the Google image results are also lacking in terms of skin-tone diversity and intersectional diversity.

We will test the performance of *QS-balanced* and *MMR-balanced* algorithms on this Occupations dataset and compare the results, in terms of diversity and accuracy, to top Google results.

4.3.2 CelebA Dataset

Another dataset we will use for evaluation is CelebA. CelebA dataset [199] is a dataset with 202599 images of celebrities, along with a number of facial attributes, such as whether the person in the image has eyeglasses or not, whether the person

is smiling or not, etc. We will use 37 of these attributes in our evaluation. One of the attributes corresponds to whether the person in the image is "Male" or not and we will use this attribute for diversity evaluation.

We divide the dataset into two parts: train and test set. The train set (containing 90% of the images) is used to train a classification model over these attributes, which is then used to compute the query similarity score The primary dataset for summarization is the test partition of the above CelebA dataset; it contains 19962 images. The 37 facial attributes will serve as the queries to the summarization algorithm and the trained classification model will be used as the blackbox query algorithm $A(q, \cdot)$.

Some of the attributes in this dataset are gender-neutral, while others seem to be gender-specific. We consider an attribute to be gender-neutral if it is commonly associated with all genders and if the dataset has a sufficient number of images from both men and women labeled with that attribute. For example, we consider the attribute "smiling" to be gender-neutral since it is associated with both men and women, and amongst images labeled as smiling in the dataset, 34% of images that are labeled as Male and 66% are labeled as Female. [9] Similarly, the attribute "eyeglasses" can be considered gender-neutral since it is also commonly associated with both men and women, and the dataset has a sufficient number of images of both men and women with eyeglasses. On the other hand, an attribute like "mustache" is usually associated with men and all images labeled with this attribute in the dataset are of men; hence we will consider it to be gender-specific. The fraction of images of women for other facial attributes is given in Section A.2.5 in the Appendix. Our primary goal for this dataset will be to ensure diversity with respect to such gender-neutral queries, but we will present our results for all the queries.

[9] Prior studies show that there is some correlation between gender and smiling for photographs taken during public occasions [86, 76]. However, summarization results should not reflect the bias of the source, i.e., when querying for a facial attribute like "smiling", which is associated with all genders, the results should be gender-diverse to present an unbiased picture.

Table 4.3: CelebA dataset - Comparison of top 50 images from all algorithms on the metrics of the fraction of gender anti-stereotypical images and accuracy. The accuracy is quantified as the fraction of images with the corresponding query attributes. The output returned by *QS-balanced* has a larger fraction of gender anti-stereotypical images than most of the other baselines. Only *AUTOLABEL* returns a perfectly balanced set; however at a larger loss of accuracy.

Algorithm		**Diversity metric** % gender anti-stereotypical	**Accuracy metric** avg. accuracy
Our algorithms	QS-balanced	0.23 (0.21)	0.88 (0.16)
	MMR-balanced	0.17 (0.22)	0.87 (0.16)
Baselines	QS	0.08 (0.21)	0.93 (0.16)
	DS	0.49 (0.12)	0.22 (0.21)
	MMR	0.14 (0.21)	0.92 (0.16)
	DET	0.13 (0.18)	0.90 (0.17)
	AUTOLABEL	0.50 (0)	0.80 (0.23)
	AUTOLABEL-RWD	0.07 (0.24)	0.93 (0.17)

4.4 Empirical Setup and Observations

We empirically evaluate the performance of *QS-balanced* and *MMR-balanced* on the Occupations and CelebA dataset. The complete implementation details are provided in Appendix A.2.2, including the blackbox query algorithm and the similarity function used for each of the datasets; we provide certain important details of the implementation here. In the case of the Occupations dataset, the query similarity is measured by quantifying similarity to a set of images corresponding to the query, while in the case of the CelebA dataset, the query similarity is measured using the output of a classifier pre-trained on the training partition of the dataset.

Since the choice of the control set is dataset and domain-dependent, we discuss the content and construction of control sets used for our simulations. A detailed discussion on the composition, social, and policy aspects of the control sets is presented in Section 4.5.

4.4.1 Control Sets

Similar to the previous chapter, the chosen control set should satisfy Assumption 3.3.1 stated in Chapter 3. That is, the control set of images should satisfy the following criteria: (a) the control set should consist of a small number of images that belong to the same domain as the dataset, and (b) the images should primarily differ with respect to the protected attribute and stay similar with respect to other attributes, such as background, face positioning, etc.

For the Occupations dataset, we evaluate our approach on four different small control sets. Two sets (with 12 images each) are hand-selected using images from Google results and are intended to be diverse with respect to presented gender and skin color. The reason for using Google search to construct these sets was simply to ensure that the set is comprised of images from the same domain as the dataset itself. These images are also not part of the Occupations dataset. The other two sets (with 24 images each) are generated by randomly sub-sampling from the Pilot Parliaments Benchmark (PPB) dataset [39]. We use the PPB dataset to construct control sets because it contains portrait images of parliamentarians from different countries, and thus ensures that the images predominantly highlight the facial features of the person. The images in the PPB dataset have gender and skin-tone labels, and we randomly select 24 images for our control set, conditioned on the sampled set containing an equal number of images of men and women and an equal number of images of different skintones. These control sets are presented in Section A.2.4.

For the CelebA dataset, once again we use four different control sets for our evaluation, two of them have 8 images and the other two have 24 images; the exact images are provided in Appendix A.2.5. The control sets are constructed by randomly sampling an equal number of images with and without the "Male" attribute from the train set. Once again, we use the training part of the dataset to

construct control sets because, if possible, the images in the control sets should be from the same domain as the dataset itself. Since the domain, in this case, is images of celebrities, using images from the training partition leads to better results (in terms of accuracy and diversity) than using images from Google search.

The results presented here compare the best performance using one of the control sets and the comparison of different control sets is presented in the Appendix.

4.4.2 Baselines

To better judge the results of our algorithms, we compare them to multiple other approaches as well as relevant baselines. We first consider two baselines that give the range of our options – simply considering query accuracy (*QS*), or simply considering the diversity of the set (*DS*). We also compare our results to the existing top Google results in the dataset. For other baselines, we consider natural and effective approaches that have been proposed in prior image summarization literature. To score images on query relevance, all algorithms once again either measure similarity using query images, in the case of the Occupations dataset, or use the output of the trained classifier, in the case of the CelebA dataset. To ensure diversity in the summary, prior work can be divided into two categories: algorithms that aim to reduce redundancy in the summary and algorithms that use protected attribute labels inferred using pre-trained classification tools. We compare both kinds of algorithms, and also discuss the potential drawbacks of these approaches below.

Algorithms that ensure non-redundancy

Reducing redundancy is a common approach for achieving diversity in the output summary. Essentially, algorithms that aim to maximize non-redundancy try to choose a summary that has images that are *maximally-representative* of all the rele-

vant images. However, as shown by prior work [51] and our empirical results, this approach does not always effectively diversify across protected attributes, such as gender, and instead results in a summary that is diverse with respect to other attributes, such as background, body position, etc. We compare our algorithms against two approaches that fall under the category of reducing redundancy in the output summary.

- *DET*: Determinant-based diversification [177, 52]. This approach first filters images according to their query relevance. Then it uses a geometric measure (determinant) on the features of a given subset of relevant images to quantify the diversity of the subset and aims to select the subset that maximizes this measure of diversity. However, without any constraints on the subset, *DET* returns a summary that is diverse across all features, including irrelevant features such as background color, and hence can be unsuitable for the task of diversifying across the given protected attributes.

- *MMR*: This algorithm is an iterative greedy algorithm that starts with an empty set and, in each iteration, adds an image that has *maximum marginal relevance*, a score that combines both query relevance and extent of similarity to the images already chosen for the summary [46]. Similar to *DET*, we compare against this method to show that greedily choosing non-redundant images does not explicitly lead to diversity across protected attribute values.

Algorithms that use label-inference tools

Many existing fair summarization algorithms assume the presence of protected attribute labels to generate fair summaries [193, 52], by using labels to enforce *fairness constraints* on the output summary. In the absence of labels, one way to employ these algorithms is to use pre-trained classification tools to infer the protected attribute labels for all images in the dataset. For example, one can use pre-trained

gender classification tools to obtain gender labels for the images and then enforce constraints using these inferred labels. However, this approach can be problematic if the classification model has been trained on biased data (as seen in [39]) or has a relatively low accuracy for the given dataset. In both cases, the use of a pre-trained gender classification model can further exacerbate the bias in the summary (as will be evident from empirical results on the Occupations dataset). For comparison of our approach against these kinds of methods, we use a pre-trained gender classification model [188] and the following two approaches for generating summaries using query similarity scores and inferred labels.

- *AUTOLABEL:* Using pre-trained gender classification model [188] [10], this approach first divides the dataset into two partitions: images labeled "male" and images labeled "female". Then it sorts images in each partition by query relevance score and selects an equal number of top images labeled "male" and "female" for the summary.

- *AUTOLABEL-RWD*: Once again using the same pre-trained gender classification model, along with a more effective scoring function suggested by [193]; this approach rewards a subset for having images from multiple partitions instead of penalizing it for having images from the same partition.

Empirical comparison with these baselines will show that the bias or errors in pre-trained classification models can often exacerbate the bias of generated summaries or adversely affect their accuracy.

Additional mathematical details and descriptions of all the baselines are provided in Section A.2.1 of the Appendix. Each algorithm, including the baselines, is used to create a summary of 50 images, corresponding to each query occupation. The comparison of our algorithms and baselines on smaller summary sizes is also

[10] https://github.com/dpressel/rude-carnie

presented in Section A.2.4 and A.2.5 in the Appendix. For the Occupations dataset, we compare our algorithm and the baselines on metrics of gender diversity, skin color diversity, and accuracy. For the CelebA dataset, we compare our algorithm and the baselines on metrics of gender diversity and accuracy.

4.4.3 Observations - Gender Diversity

Occupations dataset

As reported earlier, 52 out of 96 occupations have a larger fraction of men employed and the rest have a larger fraction of women employed (inferred using the BLS data [2]). We first report the fraction of gender anti-stereotypical images in the output for each query occupation, i.e., if an occupation is male-dominated, we take into account the fraction of women and if an occupation is female-dominated, we take into account the fraction of men in the output set. The results are presented in Table 4.2. Algorithm *QS-balanced*, using PPB Control Set-1 returns a set for which the average fraction of gender-anti-stereotypical images is 0.45 with a standard deviation of 0.17. In comparison, for Google Image search, the average fraction of gender-anti-stereotypical images in top results is 0.30 with a standard deviation of 0.22. The table shows that *QS-balanced* algorithm returns a larger fraction of images that do not correspond to the gender stereotype associated with the occupation.

In terms of raw gender numbers, the average fraction of women in top results of *QS-balanced*, for any occupation is 0.35 with a standard deviation of 0.10. The results for the performance of *QS-balanced* using other control sets are presented in Section A.2.4 of the Appendix. Using control set-1 leads to a slightly larger average fraction of women; however using PPB Control Set-1 leads to better performance with respect to both gender and skintone, which is why we present our main re-

sults using this control set.

The gender diversity of the results of *MMR-balanced* is similar to those of *QS-balanced* and much better than Google results and baselines. The average fraction of gender anti-stereotypical images in the *MMR-balanced* is 0.45, with a standard deviation of 0.20, which is slightly worse than *QS-balanced* results. The average fraction of women in top results of any occupation for *MMR-balanced* is 0.40 with a standard deviation of 0.17. The results empirically show that the use of a control set appropriately, either in *QS-balanced* or *MMR-balanced*, leads to better diversification across gender.

The variation of the percentage of women in the output of different algorithms is presented in Figure 4.4(a). The x-axis in Fig 4.4(a) is the actual percentage (ground truth) of women in occupations, obtained using data from BLS [2]. The figure primarily shows the results from *MMR-balanced* and *QS-balanced* are relatively more gender-balanced, On the other hand, *MMR* and *DET* have a relatively smaller fraction of gender anti-stereotypical images in their output. This shows that algorithms that aim to diversify across feature space (like *MMR* and *DET*) cannot always achieve desired diversity with respect to protected attributes, such as gender. The fraction of gender anti-stereotypical is however better than Google results, showing that it does diversify across gender to an extent.

The performance of gender anti-stereotypical images in the output of *AUTOLABEL* and *AUTOLABEL-RWD* is relatively low as well (around 0.35); this is likely due to the low accuracy of the auto-gender classification tool used (error rate 30%). The performance of these algorithms shows that one cannot rely on automatic classification tools, for gender or other protected attributes, to ensure constraint-based diversification. Hence, an intervention, in the form of a control set, can help target the necessary attributes appropriately.

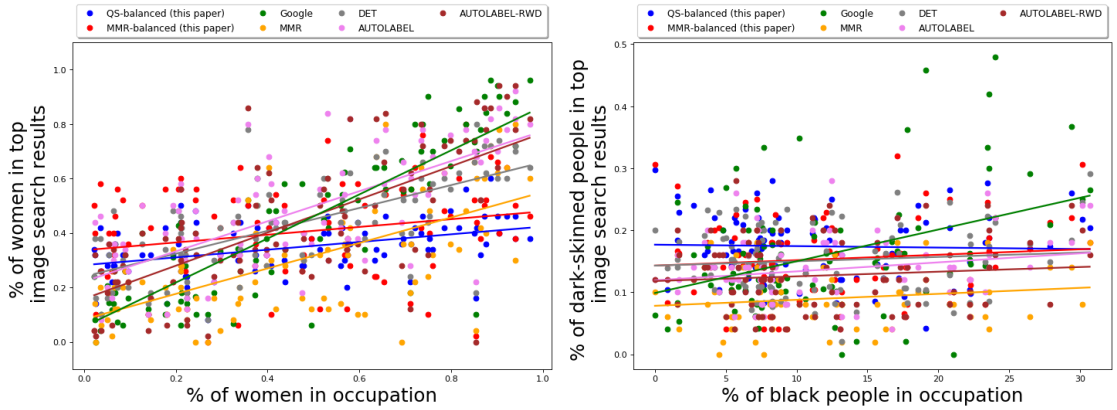
CelebA dataset

Table 4.3 shows that the output images of *QS-balanced* algorithm contain a larger fraction of gender anti-stereotypical images (0.23) than *MMR-balanced*, *MMR*, *DET*, *AUTOLABEL-RWD*. The average loss in accuracy is also small (0.05) for *QS-balanced*.

On the other hand, the output set from *AUTOLABEL* algorithm is always perfectly balanced. This is because the auto-gender classification tool used for the CelebA dataset has much better accuracy (95%), and hence we are always able to choose a perfectly gender-balanced set. However, the accuracy of this algorithm is relatively much worse than other algorithms; showing that enforcing hard fairness constraints does not always lead to the best results.

Even for image sets from *QS-balanced* and *MMR-balanced*, the overall fraction of gender anti-stereotypical images is not close to 50%, as desired. This is primarily because many queries correspond to a gender stereotype; for example, most of the images satisfying the attribute "wearing necklace" correspond to female celebrities and hence the algorithm cannot diversify with respect to this feature, due to the lack of images of men satisfying this attribute. Similarly, most of the images satisfying the attribute "bald" correspond to male celebrities, and hence the images for this query mostly contain men.

On the other hand, our framework does lead to more gender-balanced results for queries that do not have an associated gender stereotype. For example, for the query "smiling", the top 50 images with the best query scores contains only images of women, whereas the results from *QS-balanced* contain around 36% men and 64% women images. Similarly, for the query "receding hairline", the top 50 images with the best query scores contains 12% women, whereas *QS-balanced* returns an image set with 38% women. Hence, for queries that are gender-neutral, using our framework leads to results that are relatively more gender-balanced.



(a) Gender diversity comparison (b) Skin-tone diversity comparison

Figure 4.4: Occupations dataset: (a) Percentage of women in top 50 results vs ground truth of percentage of women in occupations. The images are generated using *QS-balanced*, *MMR-balanced*, and other baselines for the Occupations dataset. The figure shows that the image results from *QS-balanced* and *MMR-balanced* are more gender-balanced (see also Table 4.2), than image results from other algorithms. While the fraction of images of women from *QS-balanced* is slightly lower than *MMR-balanced*, the fraction of gender-anti-stereotypical images for both algorithms is close (see Table 4.2). (b) Percentage of dark-skinned people in top 50 results vs ground truth of percentage of Black people in occupations. The image results from *QS-balanced* are relatively more balanced with respect to skintone; however, the fraction of images of dark-skinned people is low for all algorithms.

4.4.4 Observations - Skin-tone Diversity

Occupations dataset

Unlike gender, for skintone, dark-skinned people are the minority group for all occupations considered in this dataset. Hence, in this case, the fraction of anti-stereotypical images just corresponds to the fraction of images of dark-skinned people.

Using Algorithm *QS-balanced*, with PPB Control Set-1, the average fraction of people with dark skintone in top results of any occupation is 0.17 with a standard deviation of 0.05; for Google Image search, the average fraction of women in the top 50 results for any occupation is 0.16 with a standard deviation of 0.09. The high standard deviation shows that Google results are relatively more imbalanced

with respect to gender, i.e., for many occupations, the fraction of images of dark-skinned people is much smaller or larger than the average. The skin-tone diversity of the results of *MMR-balanced* is also relatively better than baselines; the average fraction of women in top results of any occupation is 0.15 with a standard deviation of 0.06.

We also compare the skin-tone diversity of results of *QS-balanced* with other baseline algorithms; the results are presented in Table 4.2 and Figure 4.4(b). The x-axis in Fig 4.4(b) is the actual percentage (ground truth) of Black people in occupations, once again obtained using data from Bureau of Labor and Statistics [2].

Once again *MMR* is unable to diversify across the desired attributes. For the results obtained using *MMR*, the average fraction of people with dark skintone in top results is 0.09, with a standard deviation of 0.05. The skin-tone diversity of results of *DET* is relatively better, the average fraction of people with dark skintone in top results is 0.15, with a standard deviation of 0.05.

Note that for all algorithms, the top results still have a very small fraction of people with dark skintone (despite using a control set that is balanced with respect to skintone). This is primarily because, for most occupations, there are very few images of people with dark skin-tone in the dataset. We expect that summarization over a more robust dataset (such as one accessible to Google for search results) can lead to better results.

4.4.5 Intersectional Diversity

In the presence of multiple protected attributes, intersectional diversity would imply that the results are diverse with the respect to the combination of the protected attributes.

Occupations dataset

We evaluate the performance of *QS-balanced algorithm* on the basis of intersectional diversity with respect to gender and skin-tone attributes. In other words, we check how the output set is distributed across the following four partitions: gender stereotypical fair skin-tone images, gender anti-stereotypical fair skin-tone images, gender stereotypical dark skin-tone images, and gender anti-stereotypical dark skin-tone images. The results are presented in Table 4.1. The control set used here is PPB Control Set-1.

As discussed earlier, Google images tend to favor the gender and skintone associated with the stereotype of the occupation; the table shows that the fraction for gender-stereotypical fair skin-tone images is much larger than the fraction for other partitions. In comparison, the results from *QS-balanced* are relatively more balanced; the difference between the fraction of gender-stereotypical and gender anti-stereotypical images is smaller, for both fair skintone and dark skintone. Furthermore, the fraction of gender anti-stereotypical dark skin-tone images in the output of *QS-balanced* is also larger than the corresponding fraction in Google images. The comparison with other baselines is also presented in Table A.1 in the Appendix.

Overall, the fraction of gender anti-stereotypical dark-skinned images is still low in the output of *QS-balanced*. Once again, the primary reason for this is the lack of robustness of the dataset itself. As noted earlier, for 35 occupations, the dataset does not contain any gender anti-stereotypical dark-skinned images; to choose such images for these queries, the algorithm has to look for similarity with images from other occupations, which leads to a small fraction of gender anti-stereotypical dark skinned images and also affects accuracy.

4.4.6 Observations - Accuracy

Occupations dataset

For the Occupations dataset, we compute accuracy by measuring similarity to the query in the following manner: for every query occupation q, we have a small set of images T_q for reference; for example, for query "doctor", 10 images of doctors are provided. [11]. Then using $\text{sim}(\cdot,\cdot)$ function, for the reference set T_q and for each image x in summary, we can calculate the score $\text{avgSim}_{T_q}(x) := \text{avg}_{x_q \in T_q} sim(x, x_q)$. The score $\text{avgSim}_{T_q}(x)$ gives us a quantification of how similar the image I is to all other images in set T_q, and correspondingly how similar it is to query q.[12] The query similarity of different algorithms and baselines is presented in Table 4.2. [13]

From the figure, we can see that the accuracy of the top images of *QS-balanced* (0.38) and *MMR-balanced* is relatively lower than the top images of Google image search (0.48). The average accuracy of other baselines is slightly better than our primary algorithms (greater than 0.42). Hence the loss in accuracy, due to the incorporation of the diversity control matrix, is not very large.

Note that query similarity does not imply that most of the output images belong to the query occupation. There will be images from other occupations that are matched to the query occupation since multiple occupations can have similar images (for example, doctors and pharmacists, or CEOs and financial analysts). The plot presented here simply checks whether the average query scores of the output images of *QS-balanced* and *MMR-balanced* are close to the Google search re-

[11] These images are hand-verified and are not present in the primary evaluation dataset S

[12] This is similar to the ROUGE score [192] employed to measure the utility of text summaries against reference summaries and has been shown to correlate well with human judgment.

[13] For the Occupations dataset, we can also alternately define accuracy as the fraction of images in the summary that belongs to the query occupation. However, this measure is problematic since many occupations have similar-looking images, for example, "doctor" and "chemist", or "insurance sales agent" and "financial advisor". Hence, similarity with reference images is a better measure of accuracy in this case; nevertheless, we also present the accuracy with respect to query occupation in Section A.2.4 of the Appendix.

sults and other baselines. To further check the number of images in the output set that belong to the query occupation, we plot a bar graph of the number of images belonging to the query occupation and the results are presented in Figure A.14 in the Appendix.

CelebA dataset

Table 4.3 also shows the accuracy comparison of our algorithm on the CelebA dataset against baselines. Here the accuracy is measured as the fraction of images that satisfied the query facial attribute. As expected, the accuracy of the results when using *QS-balanced* (88%) is worse than the accuracy when *QS* (93%), but better than the average accuracy of *DS* (22%), *MMR-balanced* (87%) and *AUTOLABEL* (80%). The reason for the relatively lower accuracy of *MMR-balanced* is primarily because it aims to reduce non-redundancy in the summary as well.

For some queries, such as "smiling" or "eyeglasses", the loss in accuracy is small (2%), while for other queries, such as "straight hair", even though the accuracy is small (72%), the images do visually correspond to the query. For these kinds of queries, the performance of our algorithm (in terms of accuracy and diversity) seems to be as desired. For some other queries, such as "mustache" or "wearing lipstick", the use of diversity control scores with $\alpha - 0.5$ does not seem to have an impact on gender diversity (0% gender anti-stereotypical images for both). This is primarily because these queries are associated with a gender stereotype, in which case forced diversification will affect accuracy.

4.4.7 Observations - Other Diversity Metrics

We also evaluate the performance of *QS-balanced*, *MMR-balanced* and baselines with respect to other standard diversity metrics from literature, e.g. non-redundancy scores (measured using log-determinant of the kernel matrix). The details and re-

sults of this comparison are presented in Section A.2.4 in the Appendix. To state the observations briefly, the non-redundancy scores of the output generated by *DET* are observed to be better than the non-redundancy scores of other algorithms. This is expected since *DET* optimizes the determinant-metric being measured. However, as noted before, maximizing non-redundancy does not necessarily ensure diversity with respect to gender and skintone. Amongst the proposed algorithms, *MMR-balanced* has relatively better non-redundancy scores than *QS-balanced*. This is primarily because *MMR-balanced* and has a non-redundancy component already built into it (at the cost of efficiency); *QS-balanced*, on the other hand, is faster since it only aims to ensure diversity with respect to attributes represented in the control set.

4.5 Discussion, Limitations and Future Work

The algorithms presented here are prototypes that aim to improve diversity in image summarization. A crucial feature of our framework is that it is built to extend existing image summarization algorithms (represented using the blackbox $A(\cdot, \cdot)$). This is because summarization algorithms can be designed in a manner very specific to the domain; for example, Google Image search uses the metadata of the images (such as parent website, website metadata, etc) to return images that correspond to the query. Designing a new fair summarization from scratch is unreasonable, and a post-processing approach to ensuring fairness is more likely to be adopted. However, there are certain limitations to this approach which we examine in connection to potential future work in this section.

Discussion on the observations. The empirical results show that using the control set has a positive impact on the gender and skin-tone diversity of the summary, either in the form of *QS-balanced* or *MMR-balanced* algorithm. The average fraction

of gender anti-stereotypical images in the output of both algorithms is close to 0.45, for the Occupations dataset. In comparison, the average fraction of gender anti-stereotypical images in Google images is around 0.30. Even the algorithms that aim to just reduce non-redundancy, are unable to diversify across gender and skintone to the extent that *QS-balanced* or *MMR-balanced* does.

However, the results for skintone and intersectional diversity of the results of *QS-balanced* and *MMR-balanced* on the Occupations dataset is still lower than the desired level of diversity (close to the fraction in the control set). Even though this is because of the lack of images of people with darker skintone in the Occupations dataset, it will be important to empirically evaluate the performance of the framework on more robust datasets.

In the case of the CelebA dataset, while the overall average fraction of gender anti-stereotypical images is not very high (0.23), we do observe that for certain queries, the fraction of gender anti-stereotypical images is higher than those obtained using just query scores (for example, "smiling"). These queries mostly correspond to gender-neutral facial attributes, for which there are sufficient images in the dataset.

Comparison with baselines. From the performance of *DET* and *MMR*, we see that diversifying across feature space does not necessarily diversify across the protected attributes; an observation that as also made in [51]. Furthermore, imposing hard fairness constraints (such as using *AUTOLABEL* when the pre-trained gender classifier has high accuracy) is not ideal since this can lead to an undesirably high loss of accuracy. Hence control sets can serve as a medium of *soft fairness constraints*.

Control sets. While control sets, when appropriately chosen, do seem to improve the diversity of the output, the choice of the composition of the control set

is context-dependent. It is obvious that the control set images should be chosen keeping in mind the domain of the images of the dataset, to ensure that image similarity comparison is not redundant (i.e., satisfy Assumption 3.3.1).

But what should be the fraction of images of women or dark-skinned people in the control set? We observe that changing the composition of the control set changes the composition of the output similarly. We infer this by empirically evaluating the performance of *QS-balanced* algorithm for control sets with different fractions of images of minorities and observe that as the fraction increases, the representation of images of these minorities in the output set also increases. The control sets are randomly chosen from the PPB dataset. The results of this analysis are presented in Section A.2.4 of the Appendix. Hence, the composition of the control set does seem to have an impact on the composition of the output summary.

The size of the control set is intentionally kept to be very small (recall that the time complexity depends linearly on the size of the control set). Indeed it is a key advantage of our approach that it performs well even with small control sets. Larger control sets could be used, but constructing them could be considerably more difficult, especially considering that determining the control set is context-specific and could/should require input from multiple parties. Empirically, we did not observe any statistically significant advantage in using control sets of size 100-200.

There are many other context-specific and policy-related questions about the control set that cannot be answered through the above empirical analysis. Typically for an application, the range of composition of the control set should be decided after a thorough research of the user demographics and will also require input from all the affected parties/communities to ensure that there is an appropriate representation of all groups. Once the control set is created and deployed, ideally the company responsible for the application of the framework should also

provide opportunities for public audit/examination of the criteria and diversity sets to ensure transparency in the diversification process. The reason why transparency is required in the process of selection of a control set is that, just like any other fairness metric, using misrepresentative or non-diverse control sets can lead to more harm than good. Similar to the process adopted in other settings such as voting [6], it should be up to the users to decide/judge the fairness of a control set.

Choice of tradeoff parameter α. The hyper-parameter α represents the fairness-accuracy tradeoff in this algorithm. Once again, the choice is application-oriented and depends on how much loss in accuracy is acceptable to achieve the required amount of fairness in the output. We empirically evaluate the performance of *QS-balanced* and *MMR-balanced* for different α values, and the results are presented in Appendix A.2.4 and A.2.5. As expected, as α increases from 0 to 1, the fraction of gender anti-stereotypical images (for both Occupations and CelebA datasets) increases. At the same time, the similarity to the query or accuracy decreases. In our case, the figures show that a balanced choice of $\alpha = 0.5$ is reasonable.

The choice of hyper-parameters, such as control set and α value are context-dependent and we expect the use of this algorithm to be preceded by a similar thorough evaluation and analysis using different control sets with different compositions, and different α values.

Assumption of binary protected attributes. The primary evaluation of our method (both in this chapter and Chapter 3) was with respect to binary gender and skintone. This evaluation made use of labeled data where gender and skintone were often primarily treated as binary, which can be problematically restrictive [155], an inaccurate representation of the diversity in humanity with respect to gender and skintone [118], and could be used in a discriminative manner [25, 144]. The focus on binary protected attributes in this dissertation was primarily for ease of analy-

sis. Considering the fact that we need pre-labeled or crowd-labeled datasets to assess the performance of our algorithms (i.e., assessing whether the proposed label-agnostic fair summarization algorithm achieves gender and skintone diversity or not), our analysis is limited to the range of protected attributes used in existing relevant datasets (such as the PPB and CelebA datasets) or those which can be easily labeled by crowd-annotators (such as the Occupations dataset). Nevertheless, our proposed methods can potentially be used to achieve diversity with respect to broader ranges of protected attributes. Since the diversity is incorporated using the control set, the user can employ a wide variety of images that reflect the spectrum of diversity we observe offline. However, in terms of technical assessment of our methods for non-binary protected attributes, it would be important to evaluate this work in the future over datasets that are pre-labeled with broader label classes of protected attributes.

The lack of analysis and evaluation with respect to non-binary attributes is a limitation of many existing gender classification tools as well. A study conducted by Scheuerman, Paul, and Brubaker [269] showed that existing commercial facial analysis tools do not perform well for transgender individuals and are unable to infer non-binary gender, primarily because of the focus of training on recognizing gender-stereotypical facial features. Such studies further highlight the importance of not relying on the pre-defined notion of gender, as considered by existing gender classification tools.

Dependence on blackbox algorithm A. As a post-processing approach, our proposed algorithms - *QS-balanced* and *MMR-balanced* - rely crucially on the performance of the blackbox algorithm A. If the scores returned by the blackbox algorithm are inaccurate, then the resulting post-processing algorithm will also have diminished performance in terms of both accuracy and diversity. For instance, if A

returns extremely small scores for images of people from any specific group, then it is possible that adding diversity scores using the control set will only have a small marginal effect on the overall score of images from this group. In this case, using control sets may not improve the diversity of the final summary to the desired extent. Hence, it is important to assess the performance of A before employing the proposed post-processing methods.

Limitations of Occupations dataset and crowdsourcing. The Occupations dataset that we collect and curate can serve as a potential baseline for future analysis of image summarization and retrieval algorithms. However, it is important to note that this dataset was labeled using crowdsourcing, which comes with its own limitations. While the overall set of crowdworkers was sufficiently diverse with respect to gender, there was relatively less diversity in terms of reported race and location. Insufficient heterogeneity in crowdsourcing can lead to additional biases when the majority of the crowdworkers are biased or ill-informed about certain labeling tasks [125]. The frequency of such biases is usually correlated with the complexity of the labeling task. Considering that our labeling task has relatively low complexity and the fact that we provide the crowdworkers with multiple examples of correct and incorrect labels in the beginning, we expect that group bias to not significantly affect the accuracy of labels in the Occupations dataset. Furthermore, in this Chapter, these labels are simply used as a baseline to evaluate the diversity of summaries generated by our algorithms; the performance of our algorithms will not be affected by the biases of the crowdworkers here. Nevertheless, the subjectivity of crowd annotation should be kept in mind when using the Occupations dataset for future analysis.

Better implementation techniques. Despite the control sets being balanced across male/female presented genders, the results from *QS-balanced* do not match these

ratios exactly, and there is scope for improvement, perhaps with better diversity sets or similarity functions. Our current query-matching algorithm for the Occupations dataset is based only on the similarity with the query control set images and can be improved given additional information about the image. Once again, for a model similar to Google Image search, one would have access to the metadata of the image which will help better quantify query similarity or the similarity of two images. Other transfer learning techniques, like retraining a small part of a single layer of the CNN, could also be employed for better feature extraction, although we did not see any improvement in an initial approach in this direction.

Just like other aspects of our algorithms, the implementation will also be context-specific. For example, in the case of the CelebA dataset, we had a highly-accurate multi-class classifier to determine query similarity. Hence, in this case, the accuracy of the output summaries was quite high (in the range of 85% to 90%). On the other hand, for the Occupations dataset, we had to use a generic similarity measure (average similarity with query images), which cannot be expected to have the best performance for every dataset.

Evaluation in the absence of labels. Another challenge of using this approach is that it may not always be easy to evaluate its success. Its main strength – that it can diversify without needing class labels in the training data – is also an important weakness because we may not always have labeled data with which to evaluate the results. One approach would be to predict labels using, e.g., gender classification tools [188]. However, we do not recommend using predicted labels in general as such classification tools can themselves introduce biases (as seen with the baseline *AUTOLABEL* for Occupations dataset) and are currently not designed with broader label classes or non-binary gender in mind, and hence do not address the core problem. Perhaps a better approach would be to use human evaluators to

rate or define the visible diversity of the images selected by the algorithm.

The absence of labels also limits our analysis to relatively-small datasets. Real-world image datasets handled by applications like Google Search are considerably larger than the ones used in this chapter and are often handled as data streams [214, 108]. However, without protected attribute labels, the diversity of summaries for large datasets cannot be evaluated. At the same time, since the application of our framework is independent of the labels, the performance reported in this chapter should extend to larger datasets as well, and as part of future work, exploring techniques to evaluate performance on large datasets will help establish the scalability of our approach.

Community-driven application of the framework. Our work can also be seen in the light of the push towards participatory technologies in machine learning. Uninformed application of any technology that aims to ensure fairness can inadvertently cause more harm than good [196, 318, 26]. Recent studies exploring the current and future applicability of fairness interventions have correspondingly emphasized the importance of participation of all stakeholders in the design process of an application [268, 57, 222, 90]. Such a design process is especially important for summarization models since the results of these models can shape the perceptions of the users. Participatory design encourages the practitioners to engage with the users of the application to obtain valuable feedback on the possible disparate impacts of the application and ensures that there is a balanced power relation between the user and the engineer designing an application [265, 222, 115, 185].

An important aspect of our framework is that it requires community participation to ensure its success. As discussed in Section 4.5, the selection of a control set should regularly take user feedback into account to guarantee that it is sufficiently representative of the user demographics. Encouraging community participation

also ensures that the decisions regarding key aspects of the summarization framework are not entirely made by engineers. Crucially this shifts the power of the design process away from organizations and applications like Google Search and towards the users affected by the search results.

Furthermore, a crucial advantage of our framework is its post-processing nature; given any existing blackbox summarization or ranking algorithm, our framework adds a diversification component above the blackbox algorithm to ensure that the summary is fair; hence the implementation of the framework can be independent of the organization responsible for the blackbox algorithm. This advantage can be exploited in settings where the blackbox algorithm cannot be modified. For example, our framework can possibly be implemented as a browser extension or a separate web application created by a third party that uses results from Google Image Search API and maintains a control set. However, the absence of participation of the organization that designed the blackbox summarization algorithm may also not be ideal. The engineers who design the summarization algorithm would have considerably more knowledge of the domain of the datasets and can better decide the feasibility of any control set, as well as, its impact on the accuracy of the results. As discussed earlier, an inappropriately chosen control set can lead to the exacerbation of biases in the output generated by the framework, and to prevent this, one has to make sure that the control set images belong to the same domain as the dataset. Given that the users only see a fraction of the dataset at any point in time, they cannot be expected to accurately judge the feasibility of any control set. The ideal use of control sets would, therefore, need involvement and discussion from all parties. Importantly, our framework provides an opportunity for such a discussion and can help create a balanced power dynamic between the designers of search algorithms and the users of these algorithms, when deciding how well the results should represent the user demographics.

Chapter 5

Dialect Diversity in Text Summarization on Twitter

The popularity of social media has led to a centralized discussion on a variety of topics. This has encouraged the participation of people from different communities in online discussions, helping induce a more diverse and robust dialogue, and giving voice to marginalized communities [183]. Twitter, for example, receives around 500 million posts per day, with posts written in more than 50 languages[1]. Within English, Twitter sees a large number of posts from different dialects; this diversity has even encouraged linguists to use Twitter posts to study dialects, for example, to map regional dialect variation [148, 93] or to construct parsing tools for minority dialects [31, 154]. Yet, automated language tools are often unable to handle the dialect diversity in Twitter, leading to issues like disparate accuracy of language identification between posts written in African-American English (AAE) and standard English [28], or dialect-based discrepancies in abusive speech detec-

This chapter is based on a joint work with L. Elisa Celis and was published in the proceedings of the Web Conference 2021 [167]. I would like to thank Kush Varshney for early discussions on the topic of dialect diversity.

[1] `https://www.internetlivestats.com/twitter-statistics/`

tion [267, 242].

Summarization algorithms for social media platforms, like Twitter, perform the task of condensing a large number of posts into a small representative sample. They are useful because they provide users with a synopsis of long discussions on these platforms. Yet, it is important to ensure that a synopsis sufficiently represents posts written in different dialects as the dialects are representative of the participating communities. Studies have shown that the lack of representational diversity can exacerbate negative stereotypes and lead to downstream biases [166, 280, 260, 291]. Summarization algorithms, in particular, can aggravate negative stereotypes by providing a false perception of the ground truth [166]. Hence, it is crucial for automatically generated text summaries to be dialect-diverse.

This chapter further demonstrates the efficacy of the *QS-Balanced* algorithm proposed in Chapter 4 in debiasing text summaries.

Summary of the contributions. We first analyze the dialect diversity of standard summarization algorithms that represent the range of paradigms employed for extractive summarization on platforms like Twitter. This includes frequency based algorithms (TF-IDF [203], Hybrid TF-IDF [150]), graph algorithms (LexRank [104], TextRank [209]), algorithms that reduce redundancy (MMR [122], Centroid-Word2Vec [262]), and pre-trained supervised approaches (SummaRuNNer [224]). All algorithms use various structural properties of the sentences (Twitter posts, in our case) to score them on their importance. Our primary evaluation datasets are the TwitterAAE [29], the Crowdflower Gender AI, and the Claritin datasets [77]. We observe that, for random and topic-specific collections from the TwitterAAE dataset, most algorithms return summaries that under-represent the AAE dialect. Similarly, for Crowdflower AI and Claritin datasets, these algorithms often return gender-imbalanced summaries (Section 5.2).

To address the dialect bias and utilize the effectiveness of the existing summarization algorithms, we employ the *QS-Balanced* algorithm from Chapter 4 - using any summarization algorithm as a black-box, the algorithm returns a summary that is more dialect-diverse than the summary the summarization algorithm would return without intervention. As mentioned earlier, along with the blackbox algorithm, this approach needs a small dialect-diverse control set of posts as part of the input; the generated summary is diverse in a similar manner as the control set (Section 5.3). Importantly, and in contrast to existing work [77], by using similarity metrics with items in the control set, the framework bypasses the need for dialect labels in the collection of posts being summarized.

Empirically, we show that our framework improves the dialect diversity of the generated summary for all Twitter datasets and discuss the deviation of the summaries generated by our framework from those generated by the blackbox algorithms and manually-generated summaries (Section 5.4). For the Claritin dataset, we also compare the performance against the fair summarization algorithm of Dash et al. [77], which explicitly requires labels for diversification. We observe that the summaries generated by our framework are nearly gender-balanced and ROUGE scores of these summaries (measuring the similarity between the generated and reference summaries) are close to the ROUGE scores of summaries generated by Dash et al. [77]. This comparison further exhibits the effectiveness of using control sets, instead of labels, for diversification.

Text summarization on Twitter is useful for search operations; however, there may not be a singular theme associated with the posts being summarized, which makes the context of summarization in this chapter slightly different than applications where a single document is summarized into a small paragraph [250]. In other words, the objective of this chapter can be interpreted as data-subsampling with the goal of ensuring content and representational diversity.

5.1 Related Work

Bias in NLP. Recent studies have explored the presence of social biases in various language processing models. Pre-trained encoders [210, 34, 87] have been shown to exhibit gender, racial and intersectional biases [35, 44, 288, 206, 223], often leading to social biases in downstream tasks. This includes gender and racial bias in sentiment-analysis systems [172], image captioning models [143], language identification [28, 202], hate/abusive speech detection [267, 242], and speech recognition [289]. Considering the significance of these language tasks, techniques to mitigate biases in some of the above NLP applications have been proposed [32, 35, 284, 320, 321, 77]. However, dialect diversity in summaries of textual data has not been explicitly considered before, and, in the absence of dialect labels, most fair summarization approaches cannot be extended to this problem; our work aims to address both of these issues.

Text summarization algorithms. The importance of a sentence in a collection can be quantified in different ways. Algorithms such as TF-IDF [203] and Hybrid TF-IDF algorithm [150] rank sentences based on word and document frequencies. Other unsupervised algorithms, such as LexRank [104], TextRank [209], and centroid-based approaches [262, 212, 241], quantify the importance of a sentence based on how well it represents the collection. LexRank and TextRank define a graph over the posts, quantifying the edges using pairwise similarity, and score sentences based on their centrality in the graph. Along similar lines, Rossiello et al. [262] propose a centroid-based summarization method that uses compositional properties of word embeddings to quantify the similarity between sentences.

To ensure that summary a representative of the collection being summarized, prior algorithms often define *non-redundancy* as a secondary goal [193]. This includes Maximum Marginal Relevance score (MMR) [122] algorithm, Maximum

Coverage Minimum Redundant (MCMR) models [12], Determinantal Point Processes [177], and latent variable based approaches [240, 187]. The centroid-based approach of Rossiello et al. [262] also has a non-redundancy component. While adding the sentences with the highest scores to the summary, their algorithm checks for redundancy and if a candidate sentence is *very similar* to a sentence already present in the summary, it is discarded (similar to the greedy MMR approach). However, reducing redundancy has been shown to be ineffective in ensuring diversity with respect to specific attributes, such as gender or race, in other applications [51, 50]. To empirically demonstrate the ineffectiveness of non-redundancy in ensuring dialect diversity, we analyze the summaries generated by MMR [122] and Rossiello et al. [262] (implemented using Word2Vec embeddings and referred to as *Centroid-Word2Vec* for the rest of the chapter) algorithms.

We choose TextRank and Hybrid TF-IDF for our diversity analysis because they have been shown to produce better summaries (evaluated using ROUGE metrics over manually-generated summaries) for Twitter datasets than other frequency, graph, and latent variable-based approaches [150, 230]. TF-IDF and LexRank are also commonly used for Twitter datasets and serve as baselines for our analysis. The original papers for most of these text summarization algorithms focused on the evaluation on DUC or CNN/DailyMail datasets; however, the documents in these datasets correspond to news articles that are usually not considerably dialect diverse. Beyond unsupervised approaches, supervised techniques for summarization classify whether a sentence is important to the summary or not [197, 323, 224, 151, 322]. These models are trained on datasets for which summaries are available, such as news articles [145], and the models pre-trained on these datasets do not always generalize well to other domains. We will evaluate the diversity of one such pre-trained model, SummaRuNNer [224].[2] Finally, note that Twitter posts usually

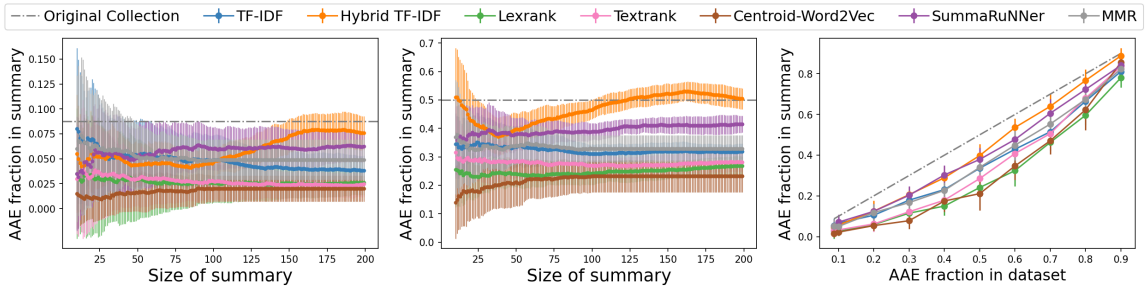
[2] *Extractive summarization* algorithms use sentences from the collection to create a summary. *Abstractive summarization*, on the other hand, aims to capture the semantic information of the dataset

have metadata associated with them, and some algorithms use this metadata to return summaries that are also diverse with respect to the time of posts [58], and/or user-network [141]. However, since our goal is to analyze the impact of dialect variation on summarization, we focus on techniques that aim to summarize using only the collection of posts.

Prior fair summarization algorithms. Most related algorithms that aim to ensure unbiased summarization usually assume the existence of labels or partitions with respect to the group attribute in consideration (in this case, dialect). For example, [52, 193] use labels to construct fairness constraints or scoring functions to guarantee appropriate diversity in automatically generated summaries. Similarly, for fair text summarization, Dash et al. [77] propose methods that use protected attribute labels to choose representative text summaries for Twitter datasets that are balanced with respect to the gender or political leaning of the users. However, these prior fair summarization approaches are unsuitable for dialect-diverse summarization since dialect labels are not always available (or even desirable [27]) for sentence collections encountered in real-world applications and automated dialect classification is a difficult task [154]. With the rapidly-evolving nature of dialects on social media, it is unreasonable to rely on existing dialect classification models to obtain accurate dialect labels for every social media post.

Using a dialect-diverse set of examples helps us skirt around the issue of unavailable dialect labels. The approach of using a diverse control set, instead of labels, to mitigate bias was employed in image-related tasks in Chapter 4, which shows that a diverse set of example images can be used to improve diversity in image summarization results and Choi et al. [62] effectively employ small refer-

and the summary creation can involve paraphrasing the sentences in the dataset [194]. Automated diversity evaluation for abstractive summarization algorithms is, therefore, more difficult since the summary is not necessarily a subset of the collection. For this chapter, we focus on extractive summarization only.



(a) 8.7% AAE posts in collection (b) 50% AAE posts in collection (c) Fixed summary size: 50

Figure 5.1: *TwitterAAE Evaluation 1.* Plots (a), (b) present the dialect diversity of generated summaries when the collection being summarized has 8.7% and 50% AAE posts respectively. Each point represents to the mean fraction of AAE posts in the summary of the given size, with standard error as errorbars. Plot (c) presents the dialect diversity in summaries of size 50 vs the original collection with varying fractions of AAE posts. All algorithms other than Hybrid TF-IDF return summaries have a smaller fraction of AAE posts than the original collection.

ence image datasets to obtain unbiased image generative models. Our framework demonstrates that such small reference sets can be used for fair text summarization as well.

5.2 Dialect Diversity of Standard Summarization Approaches

We examine the dialect diversity of TF-IDF, Hybrid TF-IDF, LexRank, TextRank, Centroid-Word2Vec, MMR, and SummaRuNNer. [3] All algorithms take as input a collection of Twitter posts and the desired summary size m, and return an m-sized summary for the collection.

[3] Algorithmic and implementation details of all methods are given in Appendix A.3.1.

5.2.1 Datasets

TwitterAAE dataset. Our primary dataset of evaluation is the large TwitterAAE dataset, curated by Blodgett et al. [29][4]. The dataset overall contains around 60 million Twitter posts from 2013, and for each post, the timestamp, user-id, and geo-location are available as well. Blodgett et al. [29] used the census data to learn demographic language models for the following population categories: non-Hispanic Whites, non-Hispanic Blacks, Hispanics, and Asians; using the learned models, they report the probability of each post being written by a user of a given population category. We pre-process the dataset to filter and remove posts for which the probability of belonging to the non-Hispanic African-American English language model or non-Hispanic White English language model is less than 0.99. This smaller dataset contains around 102k posts belonging to the non-Hispanic African-American English language model and 1.06 million posts belonging to the non-Hispanic White English language model; for simplicity, we will refer to the two groups of posts as AAE and WHE posts in the rest of the chapter.

We also isolate 35 keywords that occur in a non-trivial fraction of posts in both AAE and WHE partitions to study topic-based summarization[5]. The keywords and the fraction of AAE posts in the subset of the dataset containing them are given in Figure 5.2.

Claritin Gender dataset. Dialect variation with respect to gender has received relatively less academic attention; nevertheless, prior studies have established that there is a recognizable difference between posts by men and posts by women on Twitter [239, 213]. Hence, we look at the diversity of summarization algorithms with respect to the fraction of posts by men and women in the generated sum-

[4] `http://slanglab.cs.umass.edu/TwitterAAE`

[5] Each selected keyword occurs in at least 4500 posts in total and in at least 1500 AAE and WHE posts.

maries. The Claritin dataset contains 3943 Twitter posts about an anti-allergic drug, Claritin, with 38% from male user accounts and 62% from female user accounts[6]. It was curated to study the possible usage of crowdsourcing to detect gender-specific side-effects and, therefore, we look at the diversity of summaries with respect to the gender of the account users. For this dataset, three manually-generated summaries are also available [77] and will be used to evaluate the utility of our proposed fair summarization framework.

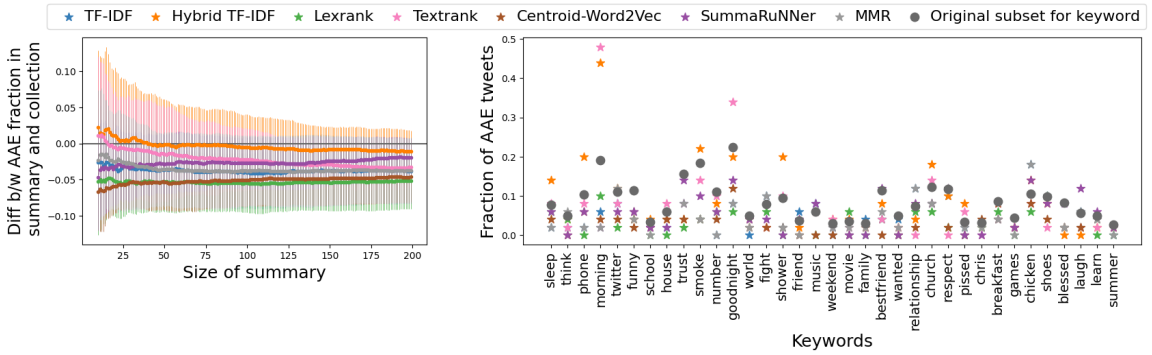
CrowdFlower AI Gender dataset. This dataset has around 20,000 posts, with crowdsourced labels for the gender of the creator of every post (male, female, or brand) and location[7]. We remove the posts with a location outside the US to maintain regional uniformity in the posts. The filtered dataset contains 6176 posts, with 34% posts from male user accounts, 35% posts from female user accounts, and the rest are labeled as posts by brands or "unknown".

For all datasets, we pre-process the posts to remove URLs, represent all posts in lower-case, replace user mentions with the tag ATMENTION and handle special characters. However, we do not remove hashtags since they are, semantically, a part of the posts.

5.2.2 Evaluation Details

Despite the filtering, the TwitterAAE dataset is prohibitively large for graph-based algorithms, due to the infeasibility of graph construction for large datasets. Hence, we limit our simulations to collections of at most size 5000 and generate summaries of sizes up to 200 for these collections.

[6] `https://github.com/ad93/FairSumm`
[7] `https://data.world/crowdflower/gender-classifier-data`



(a) Dialect diversity vs summary size (b) Dialect diversity for different keywords

Figure 5.2: *TwitterAAE Evaluation 2.* Figure (a) reports the mean and standard deviation of the difference between the AAE fraction in the summary and the AAE fraction in the collection of posts that contain the keyword. Figure (b) presents the fraction of AAE posts in size 50 summaries for different keywords, as well as, the fraction of AAE posts in the subset of posts containing the keyword. Once again, for most keywords, the algorithms (other than Hybrid TF-IDF) return summaries that have a smaller fraction of AAE posts than the original keyword-specific collection.

TwitterAAE Evaluation 1 We sample collections of 5000 posts from the TwitterAAE dataset. and vary the percentage of AAE posts in the collection from 8.7% (i.e., percentage of AAE posts in the entire dataset) to 90%. Then, we run the standard summarization algorithms for each sampled collection and record the fraction of AAE posts in the generated summaries. For each fraction, we repeat the process 50 times and report the mean and standard error of the fraction of AAE posts in the generated summaries.

TwitterAAE Evaluation 2. Next, using the 35 common keywords in this dataset, we extract the collection of posts containing any given keyword. Once again, we use the summarization algorithms on the extracted collections and report the difference between the fraction of AAE posts in the generated summary and the fraction of AAE posts in the collection containing the keyword. This evaluation aims to assess the dialect diversity of summaries generated for topic-specific sets of posts

and also lets us verify whether the observations of Evaluation 1 extend to non-random collections.

Claritin Evaluation 1. For the Claritin dataset, since the size is relatively small, we use the summarization algorithms on the entire dataset and report the fraction of posts written by men.

Crowdflower Evaluation 1. For this dataset, we again use the summarization algorithms on the entire dataset and report the fraction of posts written by men (amongst posts written by non-brands).

Remark 5.2.1. *For CrowdFlower AI and Claritin datasets, the evaluation is with respect to the gender of the user who created the post, while for the TwitterAAE dataset, the evaluation is with respect to the dialect label of the post. The evaluation methods across datasets are different in terms of the attribute used, but the goal is the same, i.e., to assess the dialect representational diversity of the generated summaries. The dialects we consider in this chapter are those adopted by social groups and the disparate treatment of these dialects is closely related to the disparate treatment of the groups using these dialects. While the AAE dialect is not necessarily used only by African-Americans, it is primarily associated with them and studies have shown disparate treatment of the AAE dialect can lead to racial bias [260, 153].*

5.2.3 Observations

The results for *TwitterAAE Evaluation 1* are presented in Figure 5.1. Plots 5.1a, b show that for small summary sizes (less than 200), all algorithms mostly return summaries that have a smaller fraction of AAE posts than the original collection. For larger summary sizes, summaries generated by Hybrid TF-IDF are relatively more dialect diverse. Even when the fraction of AAE posts in the original collec-

tion is increased beyond 0.5, the fraction of AAE posts in size 50 summaries from all algorithms is less than the fraction of AAE posts in the original collection, as evident from Figure 5.1c.

The results for *TwitterAAE Evaluation 2* are presented in Figure 5.2. For many keywords, the summaries generated by all algorithms have lower dialect diversity than the original collection. For example, for "funny" and "blessed", the AAE fraction in summaries generated by all algorithms is less than the AAE fraction in the collection containing the keyword. There are also keyword-specific collections where the summaries are relatively more diverse; e.g., for the keyword "morning", summaries generated by Hybrid TF-IDF and TextRank have better dialect diversity (AAE fraction ≥ 0.4) than the original collection (0.2). However, overall the high variance in Plot 5.2a shows that the algorithms are not guaranteed to generate sufficiently diverse summaries for all keywords.

For *Claritin Evaluation 1*, the results are presented in Table 5.1 (along with results of our "balanced" algorithms described in Section 5.3). For this dataset, all standard algorithms generate summaries that are gender-imbalanced (fraction of posts by men either ≥ 0.62 or ≤ 0.41). For *Crowdflower Evaluation 1* (Table 5.2), TF-IDF, MMR, LexRank, SummaRuNNer return nearly balanced summaries with gender fraction in the range $[0.45, 0.53]$. However, TextRank, Hybrid-TF-IDF, and Centroid-Word2Vec generate gender-imbalanced summaries (fraction of posts by men ≤ 0.37).

Discussion. The above evaluations demonstrate that none of the standard summarization algorithms consistently generate diverse and unbiased summaries across all datasets. Dialect-imbalanced original collections are not the sole reason for the dialect bias in the summaries either (as evidenced from Figure 5.1b,c). A possible reason for the bias is that the scoring mechanism of all algorithms is affected by

structural aspects of the dialect; e.g., frequency-based algorithms weigh each word in a post by its frequency. However, given that vocabulary sizes and average post lengths vary across dialects [28], using word frequency to quantify importance can favor one dialect over the other (see Section 5.5 for further discussion).

The performance of Centroid-Word2Vec and MMR for Claritin and TwitterAAE also shows that ensuring non-redundancy does necessarily not lead to dialect diversity, and the lack of diversity of SummaRuNNer summaries demonstrates that pre-trained supervised models do not necessarily generalize to other domains.

Despite the lack of dialect diversity in the generated summaries of these algorithms, prior work has demonstrated their utility [262, 250]. Hence, it is important to explore ways to exploit the utility of algorithms like Centroid-Word2Vec and, at the same time, ensure that the generated summaries are dialect-diverse.

5.3 Model to Mitigate Dialect Bias

We employ a simple framework to correct the dialect bias in standard summarization algorithms. The notations used here are similar to those in Chapter 4. Let S denote a collection of sentences. Our approach uses any standard summarization algorithm, denoted by A, as a blackbox to return a score $A(x)$, for each $x \in S$. This score represents the importance of sentence x in the collection and we assume that the larger the score, the more important is the sentence. We also need the similarity function $\text{sim}(\cdot, \cdot)$ to measure the pairwise similarity between sentences [8]. An example of such a similarity function is presented later.

To implicitly ensure dialect diversity in the results, we again use a *control set*

[8]Unlike Chapter 4, we do not use queries q as an argument for the black-box function here since we will only be summarizing data collections corresponding to a specific query or random collections. While this modification is made for simplicity of empirical analysis, one can also include the query q here if the blackbox also performs the function of finding the posts that are relevant to the given query.

T, i.e. a small set of sentences that has sufficient representation from each dialect (e.g., an equal number of posts from all relevant dialects). We return a diverse and relevant summary by appropriately combining the importance score from the blackbox A and the diversity with respect to the control set T in the following manner. Given a hyper-parameter $\alpha \in [0, 1]$, for each $z \in T$, recall the following score defined in Chapter 4:

$$\mathrm{DS}(x, x_c) = (1 - \alpha) \cdot A(x) + \alpha \cdot \mathrm{sim}(x, x_c).$$

Let DS_{x_c} represent the sorted list $\{\mathrm{DS}(x, x_c)\}_{x \in S}$ and let $DS_{x_c,i}$ denote the sentence with the i-th largest score in DS_{x_c}. Based on these scores, we rank the sentences in S in the following order: first, we return sentences that have the largest score for each x_c, i.e., $\{DS_{x_c,1}\}_{x_c \in T}$. Next, we return the set $\{DS_{x_c,2}\}_{x_c \in T}$ and so on. Sentences within each set $\{DS_{x_c,i}\}_{x_c \in T}$ can be ranked by their scores from algorithm A. At every step, for each x_c we check if a sentence has already been ranked; if so, we replace it with the sentence with the next-highest score for that x_c, ensuring that duplicates are not processed. The summary based on this ranking can then be generated. By giving equal importance to every post in T in the ranking, our framework tries to generate a summary that is diverse in a similar manner as T. This algorithm is identical to the **QS-Balanced** algorithm in Chapter 4 and the complete pseudo-code is provided in Algorithm 4. For this chapter, since we are evaluating this framework with a variety of blackbox algorithms A, we will refer to our algorithm, with blackbox A and $\alpha = 0.5$, as A-balanced. For example, our algorithm with A as Centroid-Word2Vec will be called *Centroid-Word2Vec-balanced*.

The idea of summarization based on a linear combination of scores that correspond to different goals has been used in other contexts. For topic-focused summarization, Vanderwende et al. [297] score each word by linearly adding its frequency

Table 5.1: *Claritin Evaluation 1.* We report the gender diversity and average ROUGE scores of generated summaries (size 100) against the three manually-generated summaries. For all blackbox algorithms A, our post-processed algorithm A-balanced returns more gender-balanced summaries than A (marked by •).

Method	% of posts by men in summary	ROUGE-1		ROUGE-L	
		Recall	F-score	Recall	F-score
Original collection	0.38	-	-	-	-
FairSumm	0.50	0.57	0.53	0.30	0.33
MMR	0.30	0.48	0.31	0.35	0.27
TF-IDF	0.31	0.62	0.40	0.40	0.28
TF-IDF-balanced	0.35 •	0.63	0.44	0.40	0.30
Hybrid TF-IDF	0.62	0.23	0.27	0.11	0.16
Hybrid TF-IDF-balanced	0.54 •	0.32	0.32	0.18	0.22
Lexrank	0.41	0.54	0.40	0.32	0.28
Lexrank-balanced	0.50 •	0.50	0.44	0.32	0.30
Textrank	0.62	0.22	0.24	0.09	0.14
Textrank-balanced	0.52 •	0.33	0.33	0.19	0.23
SummaRuNNer	0.35	0.62	0.49	0.42	0.32
SummaRuNNer-balanced	0.43 •	0.56	0.45	0.38	0.32
Centroid-Word2Vec	0.41	0.61	0.44	0.38	0.33
Centroid-Word2Vec-balanced	0.44 •	0.58	0.45	0.36	0.33

and topic relevance score. Even MMR computes a linear combination of the importance and non-redundancy score, measured as the maximum similarity to an existing summary sentence. As mentioned earlier, our approach is based on the fair image summarization approach used in Chapter 4 that uses diverse examples to generate a diverse image summary.

Time complexity. Let $\mathcal{T}_S$ denote the time taken by blackbox algorithm A to score all elements of S. To create the DS matrix, there will be an additive factor of $|T| \times |S|$. Selecting the best element in each DS_z can be done in two ways, i.e., either by sorting each DS_z or using a max-heap over each DS_z. In both cases, the overall time complexity is $\mathcal{T}_S + (|T| + m) \cdot |S| \cdot \log |S|$.

Choice of diversity control sets. As mentioned earlier, a diversity control set in our framework is used to ensure that generated summary has sufficient representation from every dialect. Considering the importance of the diversity control set to our framework, the appropriate construction of such sets deserves the necessary attention.

We provide one formal mechanism to construct such diversity control sets. Suppose we have a small set of dialect-labeled posts V (e.g., obtained via human annotation or crowdsourcing). To construct a control set from V, we can extract a smaller subset T (with an equal number of posts from all dialects) of V and measure how well it can predict the dialect labels of the posts in $V \setminus T$; here, the predicted label for any post $x \in V \setminus T$ is the dialect label of the post in T with which x has the highest pairwise similarity. The chosen diversity control set T is the subset with the best prediction score.

For the TwitterAAE dataset, such a V (with human-annotated dialect labels) exists [31] with $|V| = 500$. Since the time complexity of the algorithm depends linearly on the size of this set, we use the above process to select a diversity control set T of size 28 for our empirical evaluation (see Appendix A.3.2). Note that this is one way of constructing diversity control sets and, in general, the control set will be context-dependent; they can be hand chosen as well and we discuss the nuances of the composition further in Section 5.5.

5.4 Empirical Analysis of Our Model

We repeat the evaluations proposed in Section 5.2 for our post-processing framework, i.e., *TwitterAAE Evaluation 1 & 2, CrowdFlower Evaluation 1, and Claritin Evaluation 1*. For the Claritin dataset, we also compare against the FairSumm algorithm of Dash et al. [77]; FairSumm explicitly requires access to dialect labels and

Table 5.2: *Crowdflower Evaluation 1.* We report the gender diversity (fraction of non-brand posts by male user accounts) and ROUGE scores of A-balanced summaries against the summaries generated by A, for all A (summary size 100). Settings where A-balanced generates more/equally dialect-diverse summaries than A are marked with • and settings where A-balanced is worse are marked with ⋆.

Method	% of non-brand posts by men in summary	ROUGE-1		ROUGE-L	
		Recall	F-score	Recall	F-score
Original collection	0.49	-	-	-	-
MMR	0.45	-	-	-	-
TF-IDF	0.53	-	-	-	-
TF-IDF-balanced	0.44⋆	0.70	0.71	0.68	0.64
Hybrid TF-IDF	0.35	-	-	-	-
Hybrid TF-IDF-balanced	0.40 •	0.84	0.63	0.61	0.46
Lexrank	0.46	-	-	-	-
Lexrank-balanced	0.47 •	0.59	0.59	0.43	0.40
Textrank	0.37	-	-	-	-
Textrank-balanced	0.34⋆	0.82	0.81	0.78	0.73
SummaRuNNer	0.50	-	-	-	-
SummaRuNNer-balanced	0.50 •	0.76	0.73	0.66	0.68
Centroid-Word2Vec	0.34	-	-	-	-
Centroid-Word2Vec-balanced	0.40 •	0.70	0.70	0.54	0.51

comparison against this baseline lets us assess the performance of our framework, which uses diversity control sets for diversification, to an algorithm that uses attribute labels for diversification. For this dataset, Dash et al. [77] provide three manually-generated summaries of size 100 and we evaluate the summaries generated by all algorithms according to average similarity with the manually-generated summaries. The measure of evaluation employed is ROUGE recall and F-scores [192]. To state briefly, ROUGE-1 scores quantify the amount of unigram overlap between the generated summary and the reference summary, and ROUGE-L scores look at the longest co-occurring sequence in the generated and reference summary.[9] For the other datasets, since we do not have manually-generated sum-

[9]The best average ROUGE-1 recall and F-score achieved for the Claritin dataset (against the three manually-generated reference summaries), by any algorithm considered in this chapter or

maries, we use ROUGE scores to compare against summaries from the standard summarization algorithms.

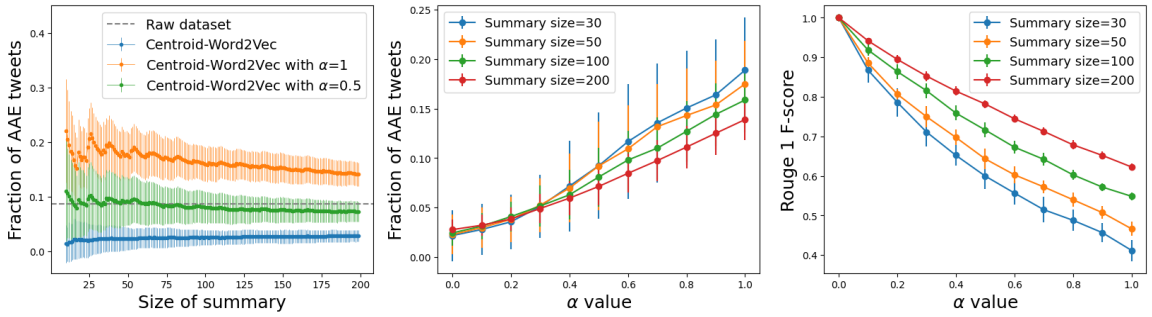
The diversity control set chosen for TwitterAAE evaluations contains 28 posts, with an equal number of AAE and WHE posts, and the sets used for Crowdflower and Claritin evaluations contain 40 and 20 posts respectively, with an equal number of posts written by male and female user accounts. Details of these sets are provided in Appendix A.3.2.

We use the following similarity function for a given pair of sentences x_1, x_2: $\text{sim}(x_1, x_2) := 1 - \text{cosine-distance}(v_{x_1}, v_{x_2})$, where v_x denotes the feature vector of sentence x. To obtain feature vectors for the sentences, we use a publicly-available word2vec model pre-trained on a corpus of 400 million Twitter posts [120]. First, we use the word2vec model to get feature vectors for the words in a sentence, and then aggregate them by computing a weighted average, where the weight assigned to a word is proportional to the smooth inverse frequency of the word (see Arora et al. [15]).

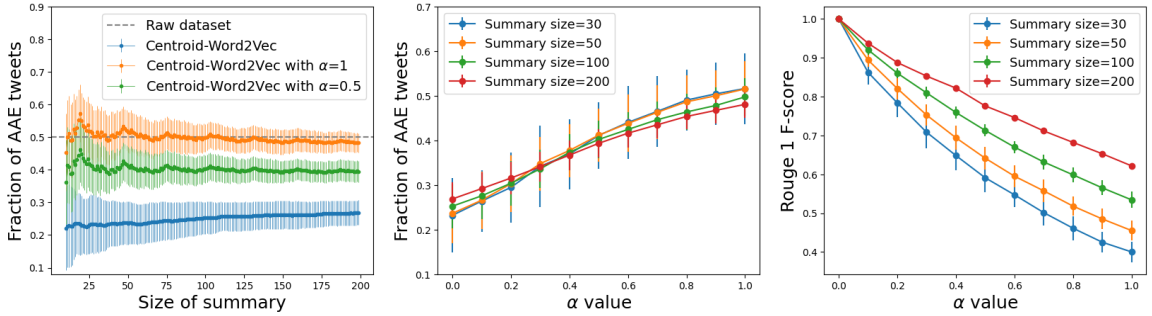
Results. The performance of our framework for *Claritin Evaluation 1* is presented in Table 5.1. We can quantify the gender balance of a summary as the deviation of the fraction of posts by men in the summary from 0.50. For all algorithms A, our framework A-balanced generates summaries that are more gender-balanced than summaries of A.

In fact, the fraction of posts by men in the summaries generated by the balanced versions of all algorithms, other than TF-IDF, is in the range $[0.43, 0.54]$. Baseline FairSumm (which requires dialect labels), as expected, returns a gender-balanced summary. ROUGE evaluation with respect to manually-generated summaries also shows that the loss in utility for some balanced algorithms, as compared to the summary generated by FairSumm, is not large. The average ROUGE-1 recall of

[77], is 0.62 and 0.57 respectively.



(a) AAE fraction vs summary size (b) AAE fraction vs α (c) Rouge-1 F-score vs α



(d) AAE fraction vs summary size (e) AAE fraction vs α (f) Rouge-1 F-score vs α

Figure 5.3: The first and second rows present the evaluation of Centroid-Word2Vec-balanced on collections containing 8.7% and 50% AAE posts respectively. Plots (a), (d) present the fraction of AAE posts for different summary sizes. Plots (b), (e) present the diversity variation with α, and plots (c), (f) present ROUGE-1 F-score between summaries generated using Centroid-Word2Vec-balanced and Centroid-Word2Vec. For both settings, Centroid-Word2Vec-balanced generates summaries that are significantly more diverse than Centroid-Word2Vec.

Centroid-Word2Vec-balanced and SummaRuNNer-balanced summaries, with respect to the three reference summaries, is 0.56 and 0.58 respectively; in comparison, the average ROUGE-1 recall of the summary generated by FairSumm is 0.57; however, the precision of Centroid-Word2Vec-balanced and SummaRuNNer-balanced summaries is slightly lower, resulting in a lower ROUGE-1 F-score compared to FairSumm summary. With respect to ROUGE-L, the Centroid-Word2Vec-balanced summary has better recall and the same F-score (0.36 and 0.33) as the FairSumm summary (0.30 and 0.33). The results show that even without access to gender labels, our framework returns nearly gender-balanced summaries, whose utility

(as measured using ROUGE evaluation with reference summaries) is comparable to that of FairSumm summary, which explicitly needs gender labels for diversification. Interestingly, for Hybrid TF-IDF and TextRank which have low initial ROUGE-1 recall (≤ 0.23) and F-scores (≤ 0.27), using our post-processing framework helps improve these utility scores by forcing the selection of a diverse set of posts. Additional manual comparison shows that reference summaries, on average, had 63 relevant posts (i.e., posts about usage or side-effects of the drug), while the summary generated by Centroid-Word2Vec-balanced had 56 relevant posts. In this context, the summary generated by our algorithm is more dialect-diverse but suffers a minimal decrease in utility.

The performance for *Crowdflower Evaluation 1* is presented in Table 5.2. Once again, the summary generated by Centroid-Word2Vec-balanced is more balanced; the fraction of non-brand posts by men in the Centroid-Word2Vec-balanced summary is 0.40, whereas it is 0.34 in Centroid-Word2Vec summary. Similarly, for Summa-RuNNer-balanced, LexRank-balanced, and TF-IDF-balanced, the fraction of posts by men in the generated summaries is in the range $[0.44, 0.50]$. However, TF-IDF-balanced and TextRank-balanced return relatively less gender-balanced summaries than their blackbox counterparts; in this case, better diversity in the summary can be achieved by using a larger α value or a different control set. The results using different α values and summary sizes are presented in Appendix A.3.4.

For *TwitterAAE Evaluation 1*, the detailed performance of our model using Centroid Word2Vec as the blackbox algorithm, is presented in Figure 5.3. Plots 5.3a,d show that using our model with $\alpha = 0.5$ (Centroid-Word2Vec-balanced) leads to improved dialect diversity in the summary (statistically different AAE fraction means). For the case when the initial collection has 50% AAE posts, Centroid-Word2Vec-balanced generates summaries that have 40% AAE posts in the summary; to achieve better dialect diversity in summary, α value needs to be increased

(Plot 5.3e). The detailed performance on *TwitterAAE Evaluation 2* for two keywords, "twitter" and "funny", is presented in Table 5.3. We see that our framework leads to a higher fraction of AAE posts in summary in most cases, compared to just the blackbox algorithm. However, it does not always improve diversity; eg, for keyword "funny" and TextRank as the blackbox, the fraction of AAE posts in summaries from the balanced version (0.04) is less than that from just the blackbox (0.06). In this case, either α or the fraction of AAE posts in the control set can be made larger to generate a more diverse summary. See Appendix A.3.3 for performance using different keywords, blackbox algorithms, and α.

The ROUGE scores for *TwitterAAE Evaluation 1* are presented in Figure 5.3c, f. As expected, the similarity between the summary generated by our model and the summary generated by Centroid-Word2vec decreases as the α increases. For summary size 200, the ROUGE-1 F-score is greater than 0.7, implying significant word overlap between the two summaries. ROUGE scores in Table 5.3 show that, for *TwitterAAE Evaluation 2*, if the diversity correction required is small, then the recall scores tend to be large. For Centroid-Word2Vec-balanced, the recall is greater than 0.64, implying that the Centroid-Word2Vec-balanced summary covers at least 64% of the words in the summary of the blackbox algorithm. However, in the cases when the summaries generated by the blackbox algorithm originally have low dialect diversity, the recall scores tend to be small (e.g., LexRank-balanced has recall around 0.5). In these cases, a larger deviation from the original summaries is necessary to ensure sufficient dialect diversity. With respect to the ROUGE assessment for TwitterAAE evaluations, note that this measure does not necessarily quantify the usability or the accuracy of the summaries in this case; it simply looks at the amount of deviation from summaries of the blackbox algorithms.

Table 5.3: *TwitterAAE Evaluation 2.* The performance of our framework for keywords "twitter" and "funny". The ROUGE scores are computed for A-balanced summaries against summaries generated by A (summary size 50). Settings where A-balanced summary has a larger fraction of AAE posts than A are marked with • and settings where A-balanced has a smaller fraction are marked with ⋆. For all but three settings, A-balanced returns summaries with a larger fraction of AAE posts than A, at the cost of certain deviation from the summaries of A.

Method	% AAE in summary	ROUGE-1 Recall	ROUGE-1 F-score	ROUGE-L Recall	ROUGE-L F-score
Keyword: "twitter"					
Collection with keyword	0.11	-	-	-	-
TF-IDF	0.10	-	-	-	-
TF-IDF-balanced	0.16 •	0.72	0.74	0.71	0.70
Hybrid-TF-IDF	0.08	-	-	-	-
Hybrid-TF-IDF-balanced	0.10 •	0.85	0.59	0.69	0.45
LexRank	0.04	-	-	-	-
LexRank-balanced	0.22 •	0.49	0.51	0.33	0.30
TextRank	0.09	-	-	-	-
TextRank-balanced	0.06⋆	0.96	0.76	0.93	0.73
SummRuNNer	0.08	-	-	-	-
SummRuNNer-balanced	0.16 •	0.57	0.55	0.42	0.40
Centroid-Word2Vec	0.06	-	-	-	-
Centroid-Word2Vec-balanced	0.12 •	0.64	0.65	0.51	0.47
Keyword: "funny"					
Collection with keyword	0.10	-	-	-	-
TF-IDF	0.04	-	-	-	-
TF-IDF-balanced	0.10 •	0.76	0.78	0.77	0.75
Hybrid-TF-IDF	0.04	-	-	-	-
Hybrid-TF-IDF-balanced	0.04⋆	0.89	0.54	0.78	0.33
LexRank	0.04	-	-	-	-
LexRank-balanced	0.22 •	0.53	0.54	0.41	0.38
TextRank	0.06	-	-	-	-
TextRank-balanced	0.04⋆	0.94	0.43	0.92	0.25
SummRuNNer	0.06	-	-	-	-
SummRuNNer-balanced	0.12 •	0.75	0.69	0.68	0.64
Centroid-Word2Vec	0.02	-	-	-	-
Centroid-Word2Vec-balanced	0.10 •	0.68	0.67	0.57	0.53

5.5 Discussion, Limitations, and Future work

Our post-processing framework provides a simple mechanism that uses standard summarization algorithms to generate diverse summaries. Yet, there are computational and societal aspects along which the framework can be further analyzed. A number of relevant socio-technical aspects of our proposed post-processing method, *QS-balanced*, are discussed in Section 4.5 of Chapter 4. This includes discussion about reliance on the performance of blackbox algorithm A, assumptions, predefined protected attributes, dependence on the choice of the control set, and community-driven implementations. In this section, we discuss other aspects of our framework that are relevant for applications of text summarization.

Analyzing the source of dialect bias. While we present empirical evidence that the standard summarization algorithms often generate dialect-biased summaries, it is critical to further delve into the source of such bias. An important empirical observation was that, for TwitterAAE evaluations, Hybrid TF-IDF generated relatively more dialect-balanced summaries than other algorithms but did not generate dialect-balanced summaries for CrowdFlower evaluation. Similarly, TF-IDF generated balanced summaries for CrowdFlower, but not for other evaluations. As mentioned earlier, this performance discrepancy of the algorithms across datasets is likely related to the design of the algorithms and the structural aspects of the posts they use to generate summaries. There are often structural differences between sentences written in different dialects. For instance, an AAE post contains around 8 words on average, while a WHE post contains around 11 words on average. The vocabulary size of all AAE posts in the TwitterAAE dataset is around 57k, while for WHE posts it is around 258k. We believe that these structural differences lead to the algorithms treating the dialects differently, resulting in dialect-imbalanced summaries. While we limit our analysis to empirical dialect diversity

evaluation, future work on this topic can explore the underlying causes for the dialect bias and suggest possible improvements to the standard summarization algorithms that directly address this bias.

Diversity control sets. While we provide an automated mechanism to construct diversity control sets (Appendix A.3.2), there are limitations to using this construction method. It crucially uses the dialect partitions in the smaller labeled dataset to construct the control set and, as discussed before, these partitions may not be desirable or capture the evolving nature of dialects. To mitigate this, the diversity control sets need to be regularly updated to include posts that better reflect the dialects of the user base.

In general, the choice of diversity control set is context-dependent, and the societal and policy impact of the control set composition requires careful deliberation. Dialects represent communities and the boundaries between dialects are quite fluid [101]. Correspondingly, deciding whether a control set sufficiently represents any specific dialect or not can be better answered by a person who writes in that dialect than by an automated classification/clustering model which constantly needs a large number of diverse sentences for training. Hence, another way to ensure that the composition of the diversity control set has sufficient representation from all user dialects is to get feedback from the communities representing the user base of the application. This would involve regular public audits and mechanisms to incorporate community assessment on the control set composition. Having a small and interpretable control set (as in our case) makes this process less cumbersome. Further, by incorporating community feedback into the design of control sets, our framework lets users have a say in the representational diversity of the summaries. Such participatory designs lead to more cooperative frameworks and are encouraged in fairness literature [268, 57].

Finally, note that using a misrepresentative control set can lead to less diverse summaries; e.g., using sentences in the control set that represent a different set of dialects than the dataset can lead to a worse summary. To prevent this, the fairness-utility tradeoff should be taken into account while deciding the control set composition.

Improved implementation. Depending on the application, the choice of pre-trained embeddings and similarity functions can be varied. For example, instead of using the cosine distance of aggregated features of all the words in a given post, one could identify words that differ across dialects and measure similarity with respect to these words only. It is also important to note that there are issues associated with ROUGE evaluations of generated summaries, such as lack of emphasis on factual correctness [175]. Recent work has proposed summary generation methods that are factually consistent [49] and extensions of our post-processing framework for such methods can be explored as part of future work.

Other domains. Another important future direction is to inspect the diversity of the algorithms for domains beyond Twitter, with sentences written in other languages, and methods to evaluate the diversity of summaries from abstractive summarization algorithms.

Chapter 6

Towards Unbiased and Accurate Deferral to Multiple Experts

Real-world applications of machine learning often involve decision-making models working together with human experts [133, 84]. For example, a model that predicts the likelihood of a disease given patient information can choose to defer the decision to a doctor who can make a relatively more accurate diagnosis [171, 253]. Similarly, risk assessment tools work together with judges and domain experts to provide a baseline recidivism risk estimate [127, 96]. Other examples of such hybrid decision-making settings include financial analysis tools [315] and content moderation tools for abusive speech detection [236] and fake news identification [282].

Human-in-the-loop frameworks are often employed in settings where automated models cannot be trusted to have high-quality inferences for all kinds of inputs. Beyond the incentive of improved overall accuracy, having human experts in the pipeline also ensures timely audits of the predictions [286] and helps fill gaps

This chapter is based on joint work with Matthew Lease and Krishnaram Kenthapadi and was published in the proceedings of AAAI/ACM Conference on AI, Ethics, and Society [170].

in the training of the automated models [243, 198]. A case in point is the study done by Chouldechova et al. [65] which showed that erroneous risk assessments by a child maltreatment hotline screening tool were frequently flagged as being incorrect by the human reviewers, implying that automated tools may not always cover the entire feature space that the domain experts use to make the decision.

However, the interaction between an ML model and a human expert is inherently more complicated than an entirely-automated pipeline. Prior studies on settings where human-in-the-loop frameworks have been implemented provide evidence of such complexities [75, 11, 119, 226]. One serious complication is the possibility of aggravated biases against protected groups, defined by attributes such as gender and race. With the increasing utilization of ML in *human classification* tasks, the problem of biases against protected groups in automated predictions has received a lot of interest. This has led to a deep exploration of social biases in popular models/datasets and ways to algorithmically mitigate them [22, 208]. Nevertheless, a number of such biased models and datasets are still in use [232]. In a pipeline that involves an interaction between a possibly-biased ML model and a human, the biases of the human can aggravate the biases of the model [248]. For example, in a study by Green and Chen [127], participants were given the demographic attributes and prior criminal record of various defendants, along with the model-predicted risk of recidivism associated with each defendant, and asked to predict the risk. They found that the participants associated a higher risk with black defendants, compared to the model prediction. In this case, the possible biases of the human in the pipeline seem to exacerbate the bias of the model prediction. Similar ethical concerns regarding the interplay between the biases of the model and humans have been highlighted in other papers [64, 254].

Motivated by the challenges discussed above, this chapter focuses on mechanisms for ensuring accuracy and fairness in hybrid machine-human pipelines. We

consider the setting where a classification model is trained to either make a decision or defer the decision to human experts. Most machine-human pipelines employed in real-world applications have multiple human experts available to share the load and to cover different kinds of input samples [65, 129]. Therefore, the hybrid decision-making framework will have an additional task of appropriately choosing one or more experts when deferring. Each expert may also have their own area of expertise as well as possible biases against certain protected groups, characterized by their prior predictions on some samples. Correspondingly, the training of a machine learning model in such a composite pipeline has to take into account the domain expertise of the humans and delegate the prediction task in an input-specific manner. Hence, our goal is to train a classifier and a deferral system such that the final predictions of the composite system are accurate and unbiased.

Summary of the contributions. We study the multiple-experts deferral setting for classification problems and present a formal *joint learning framework* that aims to simultaneously learn a classifier and a *deferrer*. The job of the deferrer is to select one or more experts (including the classifier) to make the final decision (Section 6.2.1). As part of the framework, we propose loss functions that capture the costs associated with any given classifier and deferrer. We theoretically show that, given prior predictions from the human experts and true class labels for the training samples, the proposed loss functions can be optimized using a gradient-descent algorithm to obtain an effective classifier and deferrer. Our framework further supports the settings where (a) the number of experts that can be consulted for each input is limited, (b) each expert has an individual cost of consultation, and/or (c) expert predictions are available for only a subset of training samples (Section 6.2.2). To ensure that the final predictions are unbiased with respect to a given protected attribute, we propose two fair variants of the framework (*joint*

balanced and *joint minimax-fair*) that aim to improve error rates across all protected groups. Our framework can handle both multi-class labels and non-binary protected attributes.

We empirically demonstrate the efficacy of our framework and its variants on multiple datasets: a synthetic dataset constructed to highlight the importance of simultaneously learning a classifier and a deferrer (Section 6.3.1), an offensive language dataset [80] with synthetically-generated experts (Section 6.3.2), and a real-world dataset constructed to specifically evaluate deferral frameworks with multiple available experts (Section 6.4). The real-world dataset consists of a large number of crowdsourced labels for the offensive language dataset, and is also a contribution of this dissertation. Unlike most crowdsourced datasets where the goal is simply to obtain accurate annotations, this dataset explicitly contains a dictionary of crowdworker (anonymized) to predicted labels, ensuring that the decision-making ability of each crowdworker can be inferred and consequently used to evaluate the performance of a hybrid framework like ours. We make this dataset publicly-available as this will provide a strong empirical benchmark to foster future work. For all datasets, our framework significantly improves the accuracy of the final predictions (compared to just using a classifier and other baselines, such as task allocation algorithms of Li and Liu [190] and Qiu et al. [249] from crowdsourcing literature). For the offensive language datasets, the fair variants of the framework also reduce disparity across the dialect groups.

6.1 Related Work

Given the difficulty of constructing and analyzing a human-in-the-loop framework, prior work has looked at human-in-the-loop settings from various viewpoints. One direction of research has explored the idea of the classifier having a

"reject"/"pass" option for contentious input samples [102, 191, 72, 157, 200, 71, 73]. While such an option is usually provided to ensure that low-confidence decisions can be deferred to human experts, the penalty of abstaining from making a decision in these models is fixed, and therefore, they do not take into account whether the expert at the end of the pipeline has the relevant knowledge to make the decision or not.

On the other hand, papers that take the biases and/or accuracies of the human experts into consideration are inherently more robust, but also more difficult to train and analyze. Prior theoretical models for learning to defer have constructed explicit loss functions/optimization methods to model the combined inaccuracies and biases of the classifier and the human expert [204, 218, 252, 83, 304]. Unlike the classifiers with the reject option, they use a non-static loss function for the human expert and ensure that the penalty of deferring to a human expert is input-specific. However, most of these studies primarily assume the presence of a single human expert, assuming that the expert in the pipeline will be fixed and remain the same for future classification [204, 218, 83, 304, 19]. Such an assumption is inhibitory in settings where multiple experts are available [65], as different human experts can have different prediction behaviors [130]. Raghu et al. [252] model an optimization problem for the hybrid setting as well, but they learn a classifier and a deferrer separately, which (as shown by [218] and discussed in Section 6.3) cannot handle a large variety of input settings since the classifier does not adapt to the experts. In comparison, our method learns a classifier and a deferrer simultaneously and can handle multiple experts.

Empirical studies in this direction often inherently use multiple experts since the results are based on crowdsourced data, but do not aim to propose a learning model for the pipeline [127, 319, 84, 165, 65, 165]. They, however, do highlight the importance of taking the domain knowledge of experts into account to improve

the accuracy and fairness of the entire pipeline.

Another field that studies the problem of *task allocation* among different humans is *crowdsourcing*. Crowdsourcing for data collection is a popular approach to label or curate different kinds of datasets [186]. Since crowdworkers employed for such annotation tasks come from diverse backgrounds, prior work in crowdsourcing has looked at the related issue of efficient distribution of input amongst the available workers [228, 309, 243, 298, 159, 234, 190, 249, 295]. The main difference between this line of work and our setting is the presence of the automated classifier. In our setting, the classifier is expected to handle the primary load of prediction tasks and the role of human experts is to provide assistance for input samples where the classifier cannot achieve reasonable confidence. Crowdsourcing models, however, do not usually involve the construction of any prediction model. One can alternately pre-train the classifier and treat it as another crowdworker to use task-allocation algorithms from crowdsourcing literature to distribute the samples among the experts. The main issue with this approach is that training the classifier and deferrer separately can lead to an ineffective prediction pipeline. In our empirical analysis (Section 6.3), we assess the performance of two task-allocation algorithms from crowdsourcing literature [190, 249], and demonstrate the necessity of simultaneous training. See Appendix A.4.1 for a detailed discussion on these crowdsourcing methods.

6.2 Model and Algorithms

Each sample in the domain contains a class label, denoted by $Y \in \mathcal{Y}$, n-dimensional feature vector (default attributes) of the sample used to predict the class label, denoted by $X \in \mathcal{X}$, and additional information about the sample that is available only to the experts, denoted by $W \in \mathcal{W}$. W can represent different human factors

that often assist in decision-making, such as the training or background of the expert for the given task. Let Δ_Y denote the vertices of the simplex corresponding to the unique class labels in $\mathcal{Y}$ and let $\text{conv}(\Delta_Y)$ denote the simplex and its interior. Every sample also has a protected attribute $Z \in \mathcal{Z}$ associated with it (e.g., gender or race); Z can be part of default attributes X or additional attributes W, depending on the context.

Our framework consists of a classifier and a deferrer. The classifier $F : \mathcal{X} \to \text{conv}(\Delta_Y)$, given the default attributes of an input sample, returns a probability distribution over the labels of $\mathcal{Y}$. Let $L_{\text{clf}}(F; X, Y)$ denote the convex loss associated with the prediction of classifier F at point (X, Y). For $\ell > 0$, we will call L_{clf} an ℓ-Lipschitz smooth function if for all classifiers F, $\nabla_F^2 \left(\mathbb{E}_{X,Y} L_{\text{clf}}(F; X, Y)\right) \preccurlyeq \ell \cdot \mathbf{I}$. Intuitively, Lipschitz-smoothness characterizes how fast the gradient of L_{clf} changes around any point in the parameter space of the classifier; this characterization crucially helps determine the step size required for the gradient-descent optimization of the loss function and will be useful for convergence rate bounds in our setting as well.

The framework also has access to $m - 1$ human experts $E_1, \ldots, E_{m-1} : \mathcal{X} \times \mathcal{W} \to \Delta_Y$ who can assist with the decision-making. The output of the expert will be a vector with 1 for the index of the predicted class and 0 for all other indices (one-hot encoding). The experts are assumed to have access to the additional information (from domain $\mathcal{W}$) that can be used to make the predictions more accurately; however, deferring to an expert will come at an additional cost which we will quantify later. We also assume that there is an *identity expert* which just returns the decision made by the classifier F; therefore, in total, we have m experts ($E_m(X, W) = F(X)$) (see Figure 6.1). For any given input X, the following notation

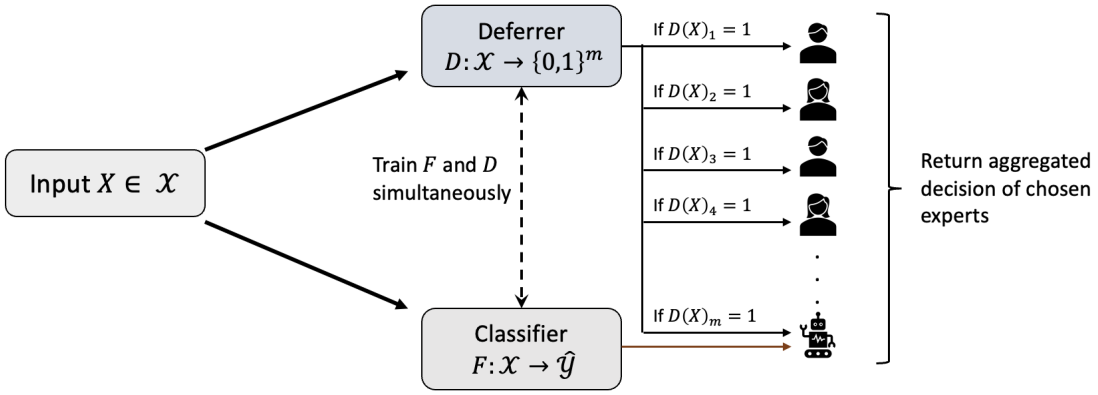


Figure 6.1: Overview of our model.

will denote all the decisions,

$$Y_E(X, W) := [E_1(X, W)^\top, \ldots, E_{m-1}(X, W)^\top, F(X)^\top].$$

The goal of the deferral system $D : \mathcal{X} \to \{0,1\}^m$, given the input, is to defer to one or more experts (including the classifier) who are likely to make an accurate decision for the given input. Given any input, D will choose a committee of experts and the final output of the framework will be based on the entries of the following matrix-vector product: $Y_E(X, W) \cdot D(X)$ (the specific aggregation method used is specified in the Section 6.2.1). If the committee chosen contains only the *identity expert*, then the output of the framework is the output of the classifier F; otherwise, the output of the model is the aggregated decision of the chosen committee.

Remark 6.2.1. *The difference between a human-in-the-loop setting and a setting with the composition of multiple prediction models [97, 36, 59] is the access to additional information W. W represents the decision-making assistance available to the experts that is not available to the prediction model either due to computational limits on the prediction model or due to lack of availability of this data for training. This assumption crucially implies that, in most cases, we cannot construct a suitably-accurate model to simulate the predictions of the experts since the importance assigned to the additional information W is*

unknown. In the absence of W, one can only try to identify the input samples for which the expert is expected to be more accurate than the trained classifier; identifying such samples using X is exactly the job of the deferrer in our framework. This distinction separates our problem setting from one where expert labels are used to bootstrap a classifier [243].

6.2.1 Simultaneously Learning Classifier & Deferrer

We first present our framework for the case of binary class label and later discuss the extension to the multi-class setting.

Binary class label, i.e, $\mathcal{Y} = \{0, 1\}$. Suppose the classifier F is fixed and, given the m experts, we need to provide a mechanism for training the deferral system (we will generalize this notion for simultaneous training shortly). For any given input X, the deferrer output $D(X)$ is expected to be a vector in the discrete domain $\{0, 1\}^m$. For the sake of smooth optimization, we will relax the domain of the output of D to include the interior of the hypercube $[0, 1]^m$, i.e., $D(X)$ will quantify the weight associated with each expert, for the given input X. Since we consider the binary class label setting, we can simplify our notation further for this section. Let $Y_{E,1}(X, W)$ denote the second row of the $2 \times m$ matrix $Y_E(X, W)$; this simplification does not lead to any loss of representational power since the sum of the first and second row is the vector $\mathbf{1}$. Along similar lines as logistic regression, using $D(X)$ one can then directly calculate the output prediction (probabilistic) as follows: $\hat{Y}_D := \sigma(D(X)^\top Y_{E,1}(X, W))$, where $\sigma(x) := e^x/(e^x + e^{1-x})$. We can then train the deferrer to optimize the standard log-loss risk function:

$$\min_D -\mathbb{E}_{X,Y}\left[Y \log(\hat{Y}_D) + (1 - Y) \log(1 - \hat{Y}_D)\right].$$

The expectation is over the underlying distribution; the empirical risk can be computed as the mean of losses over any given dataset samples (i.e., expectation over empirical distribution). For any input sample, the output prediction of the framework is 1 if $\sigma(D(X)^\top Y_{E,1}(X,W)) > 0.5$ else 0.

While the above methodology trained F and D separately, we can combine the training of the two components as well. To train F and D simultaneously, we introduce hyper-parameters α_1, α_2, and merge the loss functions for the classifier F and deferrer D linearly using these hyperparameters.

$$L(F,D) = \alpha_1 \mathbb{E}_{X,Y}\left[L_{\text{clf}}(F;X,Y)\right] - \alpha_2 \mathbb{E}_{X,Y}\left[Y \log\left(\hat{Y}_D\right) + (1-Y)\log\left(1-\hat{Y}_D\right)\right].$$

The choice of hyperparameters is context-dependent and is discussed later. The goal of the framework is then to find the classifier and deferrer pair that optimizes $\min_{F,D} L(F,D)$. We will refer to this model as the *joint framework*. The joint learning framework extends the standard logistic regression method, and hence, exhibits some desirable properties.

First, we can show that the gradient of the loss function assigns a relatively larger weight to more accurate experts.

Proposition 6.2.2 (Deferrer gradient updates)**.** *Suppose that α_1, α_2 are independent of the parameters of D. Let $Y_E \in \{0,1\}^m$ denote the decisions of the experts and classifier for any given input, and let Y denote the class label for this input. Then, for any $i \in \{1,\ldots,m\}$,*

$$-\frac{\partial L}{\partial D}^{(i)} \propto \begin{cases} e^{1-D^\top Y_{E,1}}, & \text{if } Y = 1, Y = Y_{E,1}^{(i)}, \\ -e^{D^\top Y_{E,1}}, & \text{if } Y = 0, Y \neq Y_{E,1}^{(i)}, \\ 0, & \text{otherwise.} \end{cases}$$

Here $u^{(i)}$ denotes the i-th element of vector u.

Proof. The proof of this proposition is simple. Note that

$$\sigma'(x) = \frac{2e^x e^{1-x}}{(e^x + e^{1-x})^2}.$$

Therefore,

$$\frac{\partial L}{\partial D} = -Y \cdot \frac{2e^{1-D^\top Y_{E,1}}}{e^{D^\top Y_{E,1}} + e^{1-D^\top Y_{E,1}}} \cdot Y_{E,1} + (1-Y) \cdot \frac{2e^{D^\top Y_{E,1}}}{e^{D^\top Y_{E,1}} + e^{1-D^\top Y_{E,1}}} \cdot Y_{E,1},$$

which leads to the statement of the proposition. □

The above proposition states that gradient descent moves in a direction that rewards more accurate experts. Conditional on $Y = 1$, the difference between the weight updates of a correct and an incorrect expert is proportional to $e^{1-D^\top Y_{E,1}}$. Similarly, conditional on $Y = 0$, the difference between the weight updates of a correct and an incorrect expert is proportional to $e^{D^\top Y_{E,1}}$.

Proposition 6.2.3. *$L(F, D)$ is convex in F and D, given a convex L_{clf}.*

Proof. Convexity with respect to D can be shown as an extension of the proof of Proposition 6.2.2. Taking the second derivative with respect to D also shows that it is always non-negative, implying that L is convex with respect to D. Similarly, the first part of L is convex in F (since L_{clf} is convex) and the second part contains the negative log-exponent of the product of F and the last coordinate of D, and hence is convex in F as well. □

The convexity of the function enables us to use standard gradient-descent optimization approaches [37] to optimize the loss function. In particular, we will use the projected-gradient descent algorithm, with updates of the following form:

$$F_{t+1} = F_t - \eta \cdot \left.\frac{\partial L}{\partial F}\right|_{F=F_t}, D_{t+1} = \text{proj}_{\{0,1\}^m}\left(D_t - \eta \cdot \left.\frac{\partial L}{\partial D}\right|_{D=D_t}\right),$$

where $\eta > 0$ is the learning rate and $\text{proj}_{\{0,1\}^m}(\cdot)$ operator projects a point to its closest point in the hypercube $\{0,1\}^m$. We next provide convergence bounds for the projected gradient descent algorithm in our setting when L_{clf} is Lipschitz-smooth and α_1, α_2 are constants.

Theorem 6.2.4 (Convergence bound). *Suppose L_{clf} is ℓ-Lipschitz smooth and α_1, α_2 are constants. Let $(F^\star, D^\star) := \arg\min_{F,D} L(F, D)$. Given starting point F_0, such that $\|F_0 - F^\star\| \leq \delta$, step size $\eta = c(\ell+m)^{-1}$, for an appropriate constant $c > 0$, and $\varepsilon > 0$, the projected-gradient descent algorithm, after T iterations, returns a point F°, D°, such that $L(F^\circ, D^\circ) \leq L(F^\star, D^\star) + \varepsilon$, where*

$$T = \mathcal{O}\left(\frac{(\ell + m)(\delta^2 + m)}{\varepsilon}\right).$$

Note that for $m = 1$ (just the classifier), we recover the standard gradient descent convergence bound for ℓ-Lipschitz smooth loss function L_{clf}, i.e., $O(\ell\delta^2/\varepsilon)$ iterations [37]. For $m > 1$, additionally finding the optimum deferrer results in an extra $(m(\delta^2 + \ell) + m^2)/\varepsilon$ additive term. With standard classifiers and loss functions, we can use the above theorem to get non-trivial convergence rate bounds. For example, if F is a logistic regression model and L_{clf} is the log-loss function, Lipschitz-smoothness parameter ℓ is the maximum eigenvalue of the feature covariance matrix.

To prove Theorem 6.2.4, we use the standard projected gradient-descent convergence bound stated below.

Theorem 6.2.5 ([37, 146]). *Given a convex, ℓ-Lipschitz smooth function $f : \mathbb{R}^n \to \mathbb{R}$, oracle access to its gradient, starting point $x_0 \in \mathbb{R}^n$ with $\|x_0 - x^\star\| \leq \delta$ (where $x^\star$ is an optimal solution to $\min_x f(x)$) and $\varepsilon > 0$, the projected gradient descent algorithm, with starting point x_0, step-size $\frac{1}{2\ell}$ and after $T = O\left(\frac{\ell\delta^2}{\varepsilon}\right)$ iterations, returns a point x such that $f(x) \leq f(x^\star) + \varepsilon$.*

Proof of Theorem 6.2.4. We have the following loss function:

$$L(F, D) = \alpha_1 \mathbb{E}_{X,Y}\left[L_{\text{clf}}(F; X, Y)\right] - \alpha_2 \mathbb{E}_{X,Y}\left[Y \log\big(\hat{Y}_D\big) + (1 - Y)\log\big(1 - \hat{Y}_D\big)\right].$$

The first step is to find the upper bound on the Lipschtiz-smoothness of the combined loss function. To that end, we first calculate the Lipschtiz-smoothness constants of L with respect to F and D individually. By definition,

$$\frac{\partial^2 L_{\text{clf}}}{\partial F^2} \preccurlyeq \ell I.$$

Let $L_D := \mathbb{E}_{X,Y}\left[Y \log\big(\hat{Y}_D\big) + (1 - Y)\log\big(1 - \hat{Y}_D\big)\right]$. Then,

$$\frac{\partial \hat{Y}_D}{\partial F} = 2\hat{Y}_D^2 e^{1 - 2D^\top Y_E} D^{(m)}.$$

Using the above derivative, we get that

$$\frac{\partial^2 L_D}{\partial F^2} \preccurlyeq 8e^2 \ell I.$$

Therefore,

$$\frac{\partial^2 L}{\partial F^2} \preccurlyeq (\alpha_1 + \alpha_2 8e^2)\ell I.$$

For the Lipschtiz-smoothness of L with respect to D, note that we can use results on Lipschitz-smoothness of logistic regression (since L_D corresponds to log-loss with logistic regression parameter D). In particular,

$$\frac{\partial^2 L}{\partial D^2} \preccurlyeq 2\alpha_2 \max \text{eig}(Y_E^\top Y_E) I \preccurlyeq 2\alpha_2 m I,$$

where $\max \text{eig}(\cdot)$ denotes the maximum eigenvalue of a matrix. The second inequality follows from the fact that matrix Y_E only contains 0-1 entries. For the

cross second-derivative, from proof of Theorem 6.2.2 we have that

$$\frac{\partial L}{\partial D} = -2\alpha_2 Y \cdot (1 - \sigma(D^\top Y_{E,1})) \cdot Y_{E,1} + 2\alpha_2(1 - Y) \cdot \sigma(D^\top Y_{E,1}) \cdot Y_{E,1}.$$

Therefore,

$$\frac{\partial^2 L}{\partial D \partial F} = 2\alpha_2 \sigma(D^\top Y_{E,1})^2 e^{1-2D^\top Y_{E,1}} \cdot Y_{E,1} D_m.$$

We simply need to bound the Frobenius norm of the above second derivative operator for our setting.

$$\left\| \frac{\partial^2 L}{\partial D \partial F} \right\|_F \preccurlyeq 2\alpha_2 e \sqrt{m}.$$

Therefore, combining the above inequalities, we get that the joint Lipschtiz-smoothness constant of L with respect to (F, D) (given constant α_1, α_2) is ℓ', where

$$\ell' \leq c(\ell + m),$$

where $c > 0$ is a constant. Next, since we are using a projected gradient descent algorithm, we know that the $\|D\|^2 \leq m$. Therefore, applying Theorem 6.2.5, we get that we can converge to ε-close to the optimal solution using step-size $O((\ell + m)^{-1})$ and T iterations, where

$$T = O\left(\frac{(\ell + m)(\delta^2 + m)}{\varepsilon}\right).$$

□

Our theoretical results show that, given prior predictions from the experts and true class labels for a training set, loss function L can be used to train a classifier and an effective deferrer using gradient descent.

Multi-class label. The above framework can be extended to multi-class settings as well. In this case, the matrix-vector product $Y_E(X, W) \cdot D(X)$ is a $|\mathcal{Y}|$-dimensional vector. Similar to the binary case, we extract the probability of every class label and represent it using $\hat{Y}_D$, where the j-th coordinate of $\hat{Y}_D$ represents the probability of class label being j,

$$\hat{Y}_D^{(j)} := \frac{e^{D(X)^\top Y_{E,j}(X,W)}}{\sum_{j'=1}^{|\mathcal{Y}|} e^{D(X)^\top Y_{E,j'}(X,W)}}.$$

The loss function $L(F, D)$, in this case, can be written as

$$\alpha_1 \mathbb{E}_{X,Y}\left[L_{\text{clf}}(F; X, Y)\right] - \alpha_2 \mathbb{E}_{X,Y}\left[\sum_{j=1}^{|\mathcal{Y}|} \mathbb{1}[Y{=}j] \log \hat{Y}_D^{(j)}\right].$$

The final output of the framework, for any given input, is $\arg\max \hat{Y}_D$. The above loss function retains the desired properties from the binary setting; it is convex with respect to the classifier and deferrer, and the indicator formulation ensures that each gradient step still rewards the experts that are correct for any given training input. Additional costs considered in cost-sensitive learning [325], e.g., different penalties for different incorrect predictions can also be incorporated in our framework by simply replacing the indicator function $\mathbb{1}[Y = j]$ with the penalty function [218]. For the sake of simplicity, we omit those details.

Choice of hyperparameters. α_1 and α_2 can either be kept constant or chosen in a context-dependent manner. First, note that since $\hat{Y}_D$ includes the classifier decision as well (scaled by the weight assigned to the classifier), keeping $\alpha_1 = 0$ would also ensure that the classifier and deferrer are trained simultaneously. However, due to the associated weight, classifier training with $\alpha_1 = 0$ can be slow and, since the initial classifier parameters are untrained, the classifier predictions in the initial training steps can be almost random. This will lead to the deferrer assigning a low weight to the classifier. Correspondingly, depending on the complexity of the

prediction task, it may be necessary to give the classifier a head-start as well. One way is to use time-dependent α_1, α_2. set $\alpha_1 = 1$ and $\alpha_2 = 1 - t^{-c}$, where $t \in \mathbb{Z}_+$ is the training iteration number and $c > 0$ is a constant. This choice ensures that in the initial iterations, F is trained primarily, and in the later iterations F and D are trained simultaneously.

There is a natural tradeoff associated with this head-start approach as well. The simultaneous training of F and D is crucial because the goal is to defer to experts for input where the classifier cannot make an accurate decision without the additional information. Therefore, a large head-start for the classifier can lead to a sub-optimal framework if the classifier tries to improve its accuracy over the entire domain.[1] Another choice of hyperparameters that can address this domain-partition setting is the following: set $\alpha_1{=}1$ and $\alpha_2{=}\mathbb{1}[\arg\max F(X){\neq}Y]$ so that the deferrer is trained on training samples for which the classifier is incorrect.

6.2.2 Variants of the Joint Framework

We propose several variants of the joint learning framework that are inspired by the real-world problems that a human-in-the-loop model can encounter.

Fair learning. The above joint framework aims to use the ability of the experts to ensure that the final predictions are more accurate than just the classifier. However, a possible pitfall of this approach can be that it can exacerbate the bias of the classifier, with respect to the protected attribute Z. Prior work has shown that misrepresentative training data [39, 166] or inappropriate choice of model [232], along with the biases of the human experts [127, 267] can lead to disparate performance across protected attribute types. An example of such disparity in our setting would be when, in an attempt to decrease the error rate of the prediction, the joint

[1] The synthetic experiment in Section 6.3.1 and the examples in Mozannar and Sontag [218] (for a single expert setting) highlight the necessity of simultaneously learning the classifier and deferrer.

framework assigns larger weights to the biased experts, leading to an increase in the disparity of predictions with respect to the protected attribute. We provide two approaches to handle the possible biases in our framework and ensure that the final predictions are fair.

Balanced Error Rate. One way to address the bias in final predictions is to give equal importance to all protected groups in our loss function. For protected attribute type z, let

$$\begin{aligned} L^z(F, D) :=& \alpha_1 \mathbb{E}_{X,Y|Z=z} \left[L_{\text{clf}}(F; X, Y) \right] \\ & - \alpha_2 \mathbb{E}_{X,Y|Z=z} \left[Y \log\left(\hat{Y}_D\right) + (1 - Y) \log\left(1 - \hat{Y}_D\right) \right]. \end{aligned}$$

Then the goal of this fair framework is to find the optimal solution for the problem $\min_{F,D} \sum_{z \in \mathcal{Z}} L^z(F, D)$. The above method is also equivalent to assigning group-specific weights to the samples [160, 113]. We will refer to this framework as the *joint balanced framework*.

Minimax Pareto Fairness. Martinez et al. [205]'s proposed Pareto fairness aims to reduce disparity by minimizing the worst error rate across all groups. In other words, minimax Pareto fairness proposes solving the following optimization problem: $\min_{F,D} \max_{z \in \mathcal{Z}} L^z(F, D)$.

We will employ this fairness mechanism as well and refer to this framework as the *joint minimax-fair framework*. To understand the intuition behind this framework, we theoretically show that, in the case of a binary protected attribute, the solution to the minimax Pareto fair program reduces the disparity between the risks across the protected attribute types.

Theorem 6.2.6 (Disparity of minimax-fair solution)**.** *Suppose we have a binary pro-*

tected attribute $\mathcal{Z} = \{0, 1\}$. *Let* $F^\star, D^\star := \arg\min_{F,D} \max_{z \in \mathcal{Z}} L^z(F, D)$ *denote the joint minimax-fair framework optimal solution and let* $F^\circ, D^\circ := \arg\min_{F,D} L(F, D)$ *denote the joint framework optimal solution. Then*

$$\left|L^0(F^\star, D^\star) - L^1(F^\star, D^\star)\right| \leq \left|L^0(F^\circ, D^\circ) - L^1(F^\circ, D^\circ)\right|.$$

Proof. We will first prove the theorem when
$\hat{z} := \arg\max_{z \in \mathcal{Z}} L^z(F^\star, D^\star) = 0$, i.e.,

$$L^1(F^\star, D^\star) \leq L^0(F^\star, D^\star) \leq \max_{z \in \mathcal{Z}} L^z(F^\circ, D^\circ).$$

Let $\beta = \mathbb{P}[Z = \hat{z}]$. Then for any F, D,

$$L(F, D) = \beta \cdot L^0(F, D) + (1 - \beta) \cdot L^1(F, D),$$

and by definition,

$$L(F^\star, D^\star) \geq L(F^\circ, D^\circ).$$

We will further divide the analysis into two cases. *Case 1:*

$$L^1(F^\circ, D^\circ) \leq L^0(F^\circ, D^\circ),$$

By definition of minimax-fair solution then,

$$L^0(F^\star, D^\star) \leq L^0(F^\circ, D^\circ).$$

Next, we use this inequality to look at $L^1(F^\star, D^\star)$.

$$
\begin{aligned}
& L(F^\star, D^\star) \geq L(F^\circ, D^\circ) \\
\Rightarrow & \beta \cdot L^0(F^\star, D^\star) + (1-\beta) \cdot L^1(F^\star, D^\star) \geq \beta \cdot L^0(F^\circ, D^\circ) + (1-\beta) \cdot L^1(F^\circ, D^\circ) \\
\Rightarrow & \beta \cdot L^0(F^\star, D^\star) + (1-\beta) \cdot L^1(F^\star, D^\star) \geq \beta \cdot L^0(F^\star, D^\star) + (1-\beta) \cdot L^1(F^\circ, D^\circ) \\
\Rightarrow & L^1(F^\star, D^\star) \geq L^1(F^\circ, D^\circ).
\end{aligned}
$$

Therefore, the risk disparity in this case

$$
\begin{aligned}
\left| L^0(F^\star, D^\star) - L^1(F^\star, D^\star) \right| &= L^0(F^\star, D^\star) - L^1(F^\star, D^\star) \\
&\leq L^0(F^\circ, D^\circ) - L^1(F^\circ, D^\circ).
\end{aligned}
$$

Hence the theorem is true in this case.

Case 2:

$$L^0(F^\circ, D^\circ) \leq L^1(F^\circ, D^\circ),$$

By definition of minimax-fair solution then,

$$L^0(F^\star, D^\star) \leq L^1(F^\circ, D^\circ).$$

Once again we use this inequality to look at $L^1(F^\star, D^\star)$.

$$
\begin{aligned}
& L(F^\star, D^\star) \geq L(F^\circ, D^\circ) \\
\Rightarrow & \beta \cdot L^0(F^\star, D^\star) + (1-\beta) \cdot L^1(F^\star, D^\star) \geq \beta \cdot L^0(F^\circ, D^\circ) + (1-\beta) \cdot L^1(F^\circ, D^\circ) \\
\Rightarrow & \beta \cdot L^0(F^\star, D^\star) + (1-\beta) \cdot L^1(F^\star, D^\star) \geq \beta \cdot L^0(F^\circ, D^\circ) + (1-\beta) \cdot L^0(F^\star, D^\star) \\
\Rightarrow & (1-\beta) \cdot L^1(F^\star, D^\star) \geq \beta \cdot L^0(F^\circ, D^\circ) + (1-2\beta) \cdot L^0(F^\star, D^\star) \\
\Rightarrow & (1-\beta) \cdot L^1(F^\star, D^\star) \geq \beta \cdot L^0(F^\circ, D^\circ) + (1-2\beta) \cdot L^1(F^\star, D^\star) \\
\Rightarrow & L^1(F^\star, D^\star) \geq L^0(F^\circ, D^\circ).
\end{aligned}
$$

Therefore, the risk disparity in this case

$$
\begin{aligned}
\left| L^0(F^\star, D^\star) - L^1(F^\star, D^\star) \right| = L^0(F^\star, D^\star) - L^1(F^\star, D^\star) \leq & L^1(F^\circ, D^\circ) - L^0(F^\circ, D^\circ) \\
= & \left| L^0(F^\circ, D^\circ) - L^1(F^\circ, D^\circ) \right|
\end{aligned}
$$

Hence the theorem is true in this case as well.
The proof for $\hat{z} := \arg\max_{z \in \mathcal{Z}} L^z(F^\star, D^\star) = 1$ follows by symmetry. □

Note that minimax Pareto fairness is a generalization of fairness by balancing error rate across the protected groups, but is also more difficult and costly to achieve. Furthermore, minimax Pareto fairness can handle non-binary protected attributes as well; we refer the reader to Martinez et al. [205] for further discussion on the properties of the minimax-fair solution. For our simulations, we will use the algorithm proposed by Diana et al. [88] to achieve minimax Pareto fairness.

Depending on the application, other fairness methods can also be incorporated into the framework. For example, if the fairness goal is to ensure demographic parity or equalized odds, then fairness constraints [97, 55], regularizers [162], or

post-processing methods [136, 246] can alternately be employed.

Sparse committee selection. The joint framework could assign non-zero weight to all experts. In a real-world application, requiring predictions from all of the experts can be extremely costly. To address this, we propose a sparse variant to choose a limited number of experts per input.

The number of experts consulted for any given input can be limited by using the weights from $D(X)$ to construct a small committee. Suppose we are given that the committee size can be at most k. Then, for any input X, we construct a probability distribution over the experts with probability assigned to each expert being proportional to its weight in $D(X)$, and sample k experts i.i.d. from this distribution. The final output can be obtained by replacing $D^\top Y_E$ in $\hat{Y}_D$ by the mean prediction of the committee formed by this subset (scaled by the sum of weights in D). We refer to this framework as the *joint sparse framework* when using the simple log-loss objective function, or *joint balanced/minimax-fair sparse framework*, when using an either balanced or minimax-fair log-loss objective function. We can show that the expected error disparity between joint normal and joint sparse solutions indeed depends on the properties of the distribution induced by $D(X)$.

Theorem 6.2.7 (Price of sparsity)**.** *Suppose $\mathcal{Y} = \{0, 1\}$ and let D denote the deferrer output and $\hat{Y}_D$ denote the prediction of the joint framework for a given input. Given $k \in [m]$, let random variable $\tilde{Y}_{D,k}$ denote the prediction of the joint **sparse** framework for this input. The expected difference of loss across the two predictions can be bounded as follows:*

$$\mathbb{E}\left|\log \hat{Y}_D - \log \tilde{Y}_{D,k}\right| < s_D \|D\|_1 + \max\left(2\|D\|_1, 1\right),$$

where s_D denotes the mean absolute deviation [123] of the distribution induced by D.

s_D characterizes the dispersion of the distribution induced by D and if D has low

dispersion, then the expected difference of loss from choosing a committee from distribution induced by D is low. One could also, alternately, select the experts with the k-largest weights for each input [156].

Proof of Theorem 6.2.7. Recall that in the binary class setting, given deferrer output D and expert predictions Y_E the output probabilistic prediction is calculated as

$$\hat{Y}_D := \sigma\left(D^\top Y_E\right).$$

For simplicity of presentation, since we are talking about a single input setting we are removing the input X, W in the formulas, i.e., $D(X)$ is represented as just D and $E_i(X, W)$ is just E_i. Let $E_{r_1}, \ldots, E_{r_k}$ denote the k experts sampled according to the distribution induced by $D(X)$. Then the output of the sparse framework is

$$\tilde{Y}_{D,k} := \sigma\left(\sum_{i=1}^{m} D^{(i)} \cdot \frac{1}{k} \sum_{i=1}^{k} E_{r_i}\right).$$

First, we look at $\hat{Y}_D$.

$$\log \hat{Y}_D = D^\top Y_E - \log\left(e^{D^\top Y_E} + e^{1-D^\top Y_E}\right)$$

Similarly,

$$\log \tilde{Y}_{D,k} = \sum_{i=1}^{m} D^{(i)} \cdot \frac{1}{k} \sum_{i=1}^{k} E_{r_i} - \log\left(e^{\sum_{i=1}^{m} D^{(i)} \cdot \frac{1}{k} \sum_{i=1}^{k} E_{r_i}} + e^{1-\sum_{i=1}^{m} D^{(i)} \cdot \frac{1}{k} \sum_{i=1}^{k} E_{r_i}}\right)$$

Let $N(D) := \log\left(e^{D^\top Y_E} + e^{1-D^\top Y_E}\right)$ and let

$$N'(D,k) := \log\left(e^{\sum_{i=1}^{m} D^{(i)} \frac{1}{k} \sum_{i=1}^{k} E_{r_i}} + e^{1-\sum_{i=1}^{m} D^{(i)} \frac{1}{k} \sum_{i=1}^{k} E_{r_i}}\right).$$

Then, taking the absolute difference of log-losses, we get

$$\mathbb{E}\left|\log \hat{Y}_D - \log \tilde{Y}_{D,k}\right| \leq \mathbb{E}\left|D^\top Y_E - \sum_{i=1}^{m} D^{(i)} \cdot \frac{1}{k}\sum_{i=1}^{k} E_{r_i}\right| + \mathbb{E}\left|N(D) - N'(D,k)\right|.$$

We will analyze the two terms separately. Note that for an expert sampled from distribution induced by D, we have that

$$\mathbb{E}_{r\sim D}[E_r] \cdot \sum_{i=1}^{m} D^{(i)} = D^\top Y_E.$$

Therefore,

$$\begin{aligned}\mathbb{E}\left|D^\top Y_E - \sum_{i=1}^{m} D^{(i)} \cdot \frac{1}{k}\sum_{i=1}^{k} E_{r_i}\right| &= \sum_{i=1}^{m} D^{(i)} \cdot \mathbb{E}\left|\frac{1}{k}\sum_{i=1}^{k} \mathbb{E}_{r\sim D}[E_r] - E_{r_i}\right| \\ &\leq \sum_{i=1}^{m} D^{(i)} \cdot \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\left|\mathbb{E}_{r\sim D}[E_r] - E_{r_i}\right| = \sum_{i=1}^{m} D^{(i)} \cdot s_D,\end{aligned}$$

where s_D represents the mean absolute deviation with respect to distribution induced by D. For the second absolute difference, note that both

$$D^\top Y_E, \sum_{i=1}^{m} D^{(i)} \cdot \frac{1}{k}\sum_{i=1}^{k} E_{r_i} \leq \sum_{i=1}^{m} D^{(i)}.$$

When $x > 0$,

$$\begin{aligned}\log\left(e^x + e^{1-x}\right) &= \log\left(e^{-x}(e^{2x} + e)\right) \\ &\leq \log\left(e^{2x} + e\right) \\ &\leq \log 2 + \max(2x, 1).\end{aligned}$$

Furthermore, $\log\left(e^x + e^{1-x}\right)$ is convex and achieves minimum value $0.5 + \log 2$.

Therefore, using the above upper and lower bounds, we get

$$\mathbb{E}\left|N(D) - N'(D,k)\right| \leq \max\left(2\sum_{i=1}^{m} D^{(i)}, 1\right) - 0.5.$$

Hence,

$$\mathbb{E}\left|\log \hat{Y}_D - \log \tilde{Y}_{D,k}\right| < s_D\|D\|_1 + \max\left(2\|D\|_1, 1\right).$$

□

Dropout. Given the possible disparities in the accuracies of the experts at the end of the pipeline, training a joint learning framework with diverse experts can suffer from the generalization pitfalls seen commonly in optimization literature [216]. If one expert is relatively more accurate than other experts the framework can learn to assign a relatively larger weight to this expert for every input compared to other experts. This is, however, quite undesirable as it assigns a disproportionate load to just one (or a small subset) of experts.

To tackle this issue, we introduce a random *dropout* procedure during training: an expert's prediction is randomly dropped with a probability of p and the expert's weight is not trained on the input sample for which it is dropped. This simple procedure helps reduce dependence on any single expert and ensures a relatively balanced load distribution.

Additional regularization. As mentioned earlier, the experts can have individual costs associated with their consultation. Let $C_{E_1,\ldots,E_{m-1}} : \mathcal{X} \to \mathbb{R}^{m-1}$ refer to the vector of input specific cost of each expert consultation. Assuming that the costs of the experts are independent of one another, we can take these costs into account in our framework by adding $\lambda \cdot C_{E_1,\ldots,E_{m-1}}(X)^\top D(X)_{-1}$ as a regularizer to the loss function, where $D(X)_{-1}$ denotes the first $(m-1)$ elements of the vector $D(X)$ and $\lambda > 0$ is a hyperparameter.

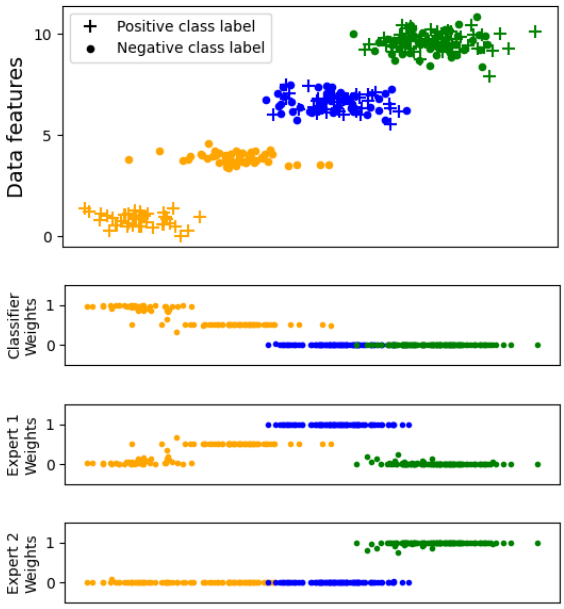


Figure 6.2: (Section 6.3.1 simulations) The first plot shows the datapoints in the synthetic dataset. The next three plots show the weights assigned to the classifier, expert 1 and expert 2 respectively for different clusters by the joint learning framework.

6.3 Synthetic Simulations

We first test the efficacy of the joint learning framework and its variants in synthetic settings. We use a synthetic and a real-world dataset for these simulations, and synthetically generate expert predictions for each input sample. For all datasets, L_{clf} will be the log-loss function and classifier F will be the standard logistic function.

6.3.1 Synthetic Dataset

Dataset and experts. Each sample in the dataset contains two features, sampled from a two-dimensional normal distribution, and a binary class label (positive or negative). There are two available experts; their behavior is described below.

Let $\mu \sim \text{Unif}(0,1)^2$ denote a randomly sampled mean vector and let $\Sigma \in \mathbb{R}^{2\times 2}$ denote a covariance matrix that is a diagonal matrix with diagonal entries sam-

pled from Unif$(0,1)$. The data has 3 clusters, represented by colors *orange*, *blue*, and *green*. The *orange* cluster has two further sub-clusters: the first sub-cluster is sampled from the distribution $\mathcal{N}(\mu, \Sigma)$ and is assigned class label 1, while the second sub-cluster is sampled from the distribution $\mathcal{N}(\mu + 3, \Sigma)$ and is assigned label 0. Since the sub-clusters are well-separated, this *orange* cluster can be accurately classified using the two dimensions.

The *blue* cluster is sampled from the distribution $\mathcal{N}(\mu + 6, \Sigma)$, and each sample is assigned a class label 1 with probability 0.5. Expert 1 is assumed to be accurate over the *blue* cluster, i.e., if a sample belongs to the *blue* cluster, expert 1 returns the correct label for that sample; otherwise, it returns a random label. Similarly, the *green* cluster is sampled from the distribution $\mathcal{N}(\mu + 9, \Sigma)$, each sample is assigned a class label 1 with probability 0.5, and Expert 2 is assumed to be accurate over the *green* cluster and random for other clusters.

We construct a dataset with 1000 samples using the above process, with an almost equal proportion of samples in each cluster; the samples are randomly divided into train and test partitions (80-20 split). The distribution of the data-points is graphically presented in Figure 6.2. Suppose the hypothesis class of classifiers is limited to linear classifiers. The ideal solution (in the absence of any expert costs) is for the classifier to accurately classify elements of the *orange* cluster, and defer the samples from *blue* cluster to expert 1 and the samples from *green* cluster to expert 2. If the linear classifier is learned before training the deferrer, then it will try to reduce error across all clusters, and the resulting framework will not be accurate over any cluster, since clusters *blue* and *green* cannot be linearly separated. By studying the performance for this synthetic dataset we can determine if the joint learning framework accurately deciphers the underlying data structure.

We also report the performance of two crowdsourcing algorithms: (a) *LL* algorithm [190] which tackles the worker selection problem, given the reliability and

Table 6.1: Overall and dialect-specific mean accuracies (standard error in brackets) for simulations in Section 6.3.2.

Method		Overall Accuracy	Non-AAE Accuracy	AAE Accuracy
Baselines	Classifier only	.89 (.00)	.86 (.00)	.96 (.00)
	Randomly selected committee	.84 (.07)	.83 (.10)	.85 (.01)
	Randomly selected fair committee	.88 (.06)	.86 (.11)	.93 (.03)
	LL	.96 (.03)	.97 (.03)	.95 (.04)
	CrowdSelect	.91 (.04)	.89 (.06)	.93 (.04)
Joint learning frameworks & fair variants	Joint framework	.92 (.02)	.89 (.03)	.97 (.00)
	Joint balanced framework	.94 (.01)	.92 (.02)	.98 (.00)
	Joint minimax-fair framework	.98 (.01)	.98 (.01)	.97 (.01)
Sparse variants of joint learning framework	Joint sparse framework	.92 (.01)	.90 (.02)	.96 (.01)
	Joint balanced and sparse framework	.92 (.01)	.89 (.01)	.97 (.00)
	Joint minimax-fair and sparse framework	.98 (.01)	.97 (.01)	.98 (.00)

variance of all the workers, and (b) *CrowdSelect* [249], which aims to model the behavior of the workers to appropriately allocate a subset of workers to each task. For both crowdsourcing algorithms, the classifier is pre-trained using the train partition and treated as just another worker. The details of these algorithms are provided in Appendix A.4.1.

Implementation details. We use projected gradient descent, with 3000 iterations, learning rate $\eta = 0.05$, and $\alpha_1 = 0, \alpha_2 = 1$. As discussed before, $\alpha_1 = 0$ can also train the classifier and deferrer simultaneously.

Results. A baseline SVM classifier trained over the entire dataset has an accuracy of around 0.67 (accurate for one cluster and random over the other two). In comparison, the joint learning framework has perfect (1.0) accuracy. If the sparse variant of the joint learning framework is used with k=1 (defer to a single expert), the accuracy drops to 0.91. To better understand the performance of the framework, Figure 6.2 presents the weights (normalized) assigned to the different experts (and classifier) for the test partition (bottom three plots).

Starting with the *green* cluster, the lowest plot shows that expert 2 is assigned

the highest weight for samples in this cluster, implying that the prediction for this cluster is always correctly deferred to expert 2. Similarly, the prediction for the *blue* cluster is always correctly deferred to expert 1. For most of the samples in the *orange* cluster, the weight assigned to the classifier is larger than the weights assigned to the two experts. For some samples in this cluster, however, a non-trivial weight is also assigned to expert 1, which is why the accuracy of the sparse variant is lower than the accuracy of the non-sparse variant. This can be prevented using non-zero expert costs, which we employ in the next simulation.

The baseline *LL* algorithm achieves an accuracy of 67% on this dataset; this is because it associates a single measure of aggregated reliability with each worker, which in this case is unsuitable since each worker has their specific domain of expertise. The *CrowdSelect* algorithm achieves the best accuracy of around 83%; in this case, the error models for each expert and the classifier are constructed individually. Due to this, the algorithm is unable to perfectly stratify the input space amongst the experts (and classifier).

Discussion. The purpose of this simulation was to show that the deferrer can choose experts in an input-specific manner. The results show that the deferrer can indeed decipher the underlying structure of the dataset, and accordingly choose the expert(s) to defer to for each input (addressing the drawback of *LL*). The important aspect of the problem to notice here is that the cluster identity is the additional information available only to the experts. The cluster identity is crucial for the experts as it reflects their domain of expertise and helps them make the correct prediction if the sample lies in their domain. On the other hand, the cluster identity is useful to the deferrer only to defer correctly; even if the cluster is part of the input, the framework cannot use it to make a correct prediction but can use it to defer to the correct expert. In other words, the framework can use the available

information to identify samples that need to be deferred to an expert (addressing the drawback of *CrowdSelect*). This sub-problem of directly identifying contentious input samples is also related to prior work by Raghu et al. [253].

6.3.2 Offensive Language Dataset

Dataset. Our base dataset consists of around 25k Twitter posts curated by Davidson et al. [80]; all posts are annotated with a label that corresponds to whether they contain hate speech, offensive language, or neither. We set the class label to 1 if the post contains hate speech or offensive language, and 0 otherwise. Using the dialect identification model of Blodgett et al. [30], we also label the dialect of the posts: African-American English (AAE) or not. Around 36% of the posts in the dataset are labeled as AAE. We treat dialect as the protected attribute in this case.

Experts. The experts are constructed to be biased against one of the dialects. We generate m synthetic experts, with $\lfloor 3m/4 \rfloor$ experts biased against AAE dialect and $\lceil m/4 \rceil$ experts biased against non-AAE dialect. To simulate the first $\lfloor 3m/4 \rfloor$ experts, for each expert $i \in \{1, \ldots, \lfloor 3m/4 \rfloor\}$, we sample two quantities: $p_i \sim \text{Unif}(0.6, 1)$ and $q_i \sim \text{Unif}(0.6, p_i)$. For expert i, p_i will be its accuracy for the non-AAE group and q_i will be its accuracy for the AAE group. To make a decision, if the input belongs to the non-AAE group then this expert outputs the correct label with probability p_i and if the input belongs to the AAE group then this expert outputs the correct label with probability q_i. By design, the first $\lfloor 3m/4 \rfloor$ experts can have a certain level of bias against the AAE group since $q_i < p_i$ for all $i \in \{1, \ldots, \lfloor 3m/4 \rfloor\}$. The same process, with flipped p_i and q_i, is repeated for the remaining $\lceil m/4 \rceil$ experts so that they are biased against the non-AAE group.

Baselines. There are three simple baselines that can be easily implemented: (1) using the classifier only, (2) randomly selected committee - a committee of size $\lceil m/4 \rceil$ is randomly selected (in this case, the predictions are expected to be biased against the AAE dialect since most of the experts are biased against the AAE dialect - see Section A.4.2), and (3) random fair committee - i.e., if the post is in AAE dialect, the committee randomly selects from experts with higher accuracy for AAE group, and if the post is in non-AAE dialect, the committee randomly selects from experts with higher accuracy for the non-AAE group. This committee selection should ensure relatively balanced accuracy across the dialects, and can therefore be used to judge the fairness of the joint learning framework. We also implement and report the performance of *LL* and *CrowdSelect* algorithms for this dataset.

Implementation details. The dataset is split into train and test partitions (80-20 split). For both classifier and deferrer, we use a simple two-layer neural network, that takes as input a 100-dimensional vector corresponding to a given Twitter post (obtained using pre-trained GloVe embeddings [244]). The experts are given a cost of 1 each, i.e., $C_{E_1,\ldots,E_{m-1}} = \mathbf{1}$ and $\lambda = 0.05$ (the regularizer used is $\lambda \cdot \mathbb{E}[C_{E_1,\ldots,E_{m-1}}(X)^\top D(X)_{-1}]$). Inspired by prior work on adaptive learning rate [95], exponent c of parameter α is set at 0.5 and dropout rate at 0.2. We present the results for $m = 20$ in this section and discuss the performance for different m, λ, and dropout rates in Appendix A.4.2. We use stochastic gradient descent for training with learning rate $\eta = 0.1$ and for 100 iterations with a batch size of 200 per iteration. For the sparse variants with $m = 20$, we sample $k = 5$ experts from the output distribution. The process is repeated 100 times, with a new set of experts sampled every time, and we report the mean and standard error of the overall and dialect-specific accuracies.

Results. The results for the joint learning framework and its variants, along with the baselines are presented in Table 6.1. The joint learning framework has a larger overall and group-specific average accuracy than the classifier. The best group-specific and overall accuracy is achieved by the joint minimax-fair framework (and its sparse variant), showing that it is indeed desirable to enforce minimax-fairness in this setting as it leads to an overall improved performance across all groups. The sparse variations of all joint frameworks, as expected, still have better performance than the classifier and random-selection baselines, and are quite similar to the non-sparse variants. Joint fair (balanced and minimax-fair) frameworks also have similar or lower accuracy disparity across the groups than random fair committee baseline. This shows that the learned deferrer is also able to differentiate between biased and unbiased experts to an extent. Due to the non-zero λ parameter used, on average, the classifier is assigned around 5% of the deferrer weight per input sample. This implies that, when creating sparse committees with $k = 5$, the classifier is consulted for around 25% of the input samples. This fraction can be further increased by appropriately increasing λ.

Further, due to our use of dropout, more accurate experts are not assigned disproportionately high weights, exhibiting the effectiveness of load balancing using dropout. This is demonstrated in Figure A.45 in the Appendix, which presents variations of the weights assigned by the joint framework to the experts vs the accuracies of the experts for a single repetition.

The *LL* algorithm is able to achieve very high overall accuracy ($\geq$ 95% for both groups) for this setting. However, our joint minimax-fair sparse framework has two advantages over *LL* algorithm. First, it achieves relatively better accuracy for both dialect groups. Second, *LL* pre-selects the most accurate experts to whom all the inputs are deferred. This is problematic and inefficient since *LL* only uses k out of m experts; in comparison, our algorithm distributes the input samples amongst

all experts to reduce the load on the most accurate experts (see Figure A.45 in Appendix). *CrowdSelect*, on the other hand, achieves lower overall and group-specific accuracies than joint minimax-fair frameworks.

6.4 Simulations Using a Real-world Offensive Language Dataset

The simulations in the previous sections highlighted the effectiveness of the joint learning framework in improving the accuracy and fairness of the final prediction. In this section, we present the results on a similar real-world dataset of Twitter posts, annotated using Mechanical Turk (MTurk).

Dataset. We use a dataset of 1471 Twitter posts for the MTurk survey. This is a subset of the larger dataset by Davidson et al. [80]. Importantly, this dataset is jointly balanced across the class categories used in Davidson et al. [80] and the two dialect groups (as predicted using Blodgett et al. [30]). Once again, the labels from Davidson et al. [80] are treated as the *gold labels* for this dataset.

MTurk experiment design. The MTurk survey presented to each participant started with an optional demographic survey. This was followed by 50 questions; each question contained a Twitter post from the dataset and asked the participant to choose one of the following options: 'Post contains threats or insults to a certain group', 'Post contains threats or insults to an individual', 'Post contains other kinds of threats or insults, such as to an organization or event', 'Post contains profanity', 'Post does not contain threats, insults, or profanity'. The options presented to the user are along the lines of the taxonomy of offensive speech suggested by Zampieri et al. [313]. The first four options correspond to offensive language in the Twitter

Table 6.2: Results of the joint learning framework and fair variants on the MTurk dataset.

Method	Overall Accuracy	Non-AAE Accuracy	AAE Accuracy
Classifier only	.78 (.02)	.76 (.05)	.80 (.04)
Joint framework	.85 (.03)	.87 (.04)	.83 (.03)
Joint balanced framework	.84 (.03)	.87 (.03)	.81 (.04)
Joint minimax framework	.85 (.02)	.87 (.02)	.83 (.02)

post, while the last option corresponds to the post being non-offensive. As in the synthetic simulations, the participants are also provided with the predicted dialect label of the post. The participants were paid a sum of $4 for completing the survey (at an hourly rate of $16).

MTurk experiment results. Overall, 170 MTurk workers participated in the survey and each post in the dataset was labeled by around 10 different annotators. Since each participant only labels a fraction of the dataset, we will treat this setting as one where there are missing expert predictions during the training of the joint learning framework. The inter-rater agreement, as measured using Krippendorff's α measure, is 0.27. As per heuristic interpretation [131], this level of interrater agreement is considered quite low for a standard dataset annotation task. However, it is suitable for our purpose since our framework aims to address situations where there is considerable disparity in the performances of different humans in the pipeline, and the goal of the joint learning framework is to choose the annotators that are expected to be accurate for the given input.

The overall accuracy of the aggregated responses (i.e., taking a majority of all responses for every post and comparing to the *gold label*) is around 87%, which is close to the accuracy of the automated classifier in Section 6.3.2 (84% for AAE posts and 91% for non-AAE posts). The high accuracy shows that using crowdsourced annotations in this setting is quite effective and the hypothetical *aggregated*

crowd annotator can indeed be considered an *expert* for this content moderation task. However, the individual accuracies of the experts is arguably more interesting and relevant to our setting.

The average individual accuracy of a participant is 77% ($\pm$13%). The minimum individual accuracy is $\approx$ 38% while the maximum individual accuracy is 98%. The wide range of accuracies evidences large variation in annotator expertise for this task. The individual accuracies for posts from different dialects also present a similar picture. The average individual accuracy of a participant for the AAE dialect posts is 76% ($\pm$15%) and the average individual accuracy of a participant for the non-AAE dialect posts is 78% ($\pm$14%).

While mean individual accuracies for the two dialects are quite similar, most annotators do display a disparity in their accuracy across the two groups. 92 of the 170 participants had a higher accuracy when labeling posts written in a non-AAE dialect. The average difference between the accuracy for non-AAE dialect posts and AAE dialect posts for this group of participants was 8.5% ($\pm$6.6%). 75 participants had a higher accuracy when labeling posts written in the AAE dialect. The average difference between the accuracy for AAE dialect posts and non-AAE dialect posts was 7.1% ($\pm$5.5%). The three remaining participants were equally accurate for both groups. The disparate accuracies here are quite similar to those in the early synthetic simulations. We next analyze the performance of the joint learning framework on this dataset.

Joint learning framework results on MTurk dataset. We perform five-fold cross-validation on the collected dataset. For each fold, we train our joint learning framework (with $\eta = 0.3$) on the train split and evaluate it on the test split. Since expert decisions are available only for a subset of the dataset, we do not use dropout or expert costs. Results are shown in Table 6.2. As before, the overall accuracy of

the joint learning frameworks is higher than the accuracy of the classifier alone. Amongst the fair variants, even though the accuracy for both dialect groups is larger when using the balanced or minimax loss function (compared to the classifier alone), it does not lead to significantly different group-specific accuracies vs. simple joint learning framework. The performance of sparse variants is presented in Appendix A.4.3. Since a relatively small number of prior predictions is available for each expert, the task of differentiating between experts here is tougher. Hence, sparse variants perform similarly or better than the classifier when committee size k is around 60 or greater.

6.5 Discussion, Limitations, and Future Work

Our proposed framework addresses settings that involve active human-machine collaboration. Having shown its efficacy for synthetic and real-world datasets, we next highlight certain limitations and fruitful directions for future work.

Fairness of the framework. It is crucial that the framework is fair with respect to the protected attribute. We proposed two methods for ensuring that the predictions are unbiased: by trying to achieve a balanced error rate for all groups, or by trying to minimize the maximum group-specific error rate (minimax Pareto fairness). Both fairness mechanisms can handle multi-class protected attributes, which helps generalize our framework to settings beyond simple binary protected attributes (e.g., multiple racial categories). An additional advantage of using these fairness definitions is that the protected group labels are not required for test or future samples, addressing the issue of their possible unavailability due to policy or privacy restrictions [99].

As mentioned in Section 6.2.2, other fairness mechanisms can also be incorporated into our framework. For most applications, the choice of fairness mechanism

and constraint is often a context-dependent question. An uninformed choice of these variables can possibly lead to a degradation of both accuracy and fairness [195] and, therefore, it is important to take the impact of any fairness constraint on the user population into account before its implementation. Similarly, in our setting, it is important to first decide whether the goal of fairness is minimizing the worst group error or demographic parity and then choose the mechanism to implement it.

Diversity of the expert pool. The wide range in accuracy observed across annotators in Section 6.4 confirms the expectation that different humans-in-the-loop will naturally bring varying levels and domains of expertise. Their accuracy will be affected by not only the training they receive but also by their background. For example, native speakers of a given dialect are naturally expected to be better annotators of language examples from that dialect. However, despite the difficulty of the task and the disparity in group accuracies, our joint learning framework is still able to identify the combination of experts that are suitable for any given input and, correspondingly, increase the accuracy and fairness of the final prediction.

Both synthetic and real-world simulations demonstrate the importance of diversity in the expert pool to achieve high predictive performance for all kinds of inputs. Human prejudices can take different forms than the biases present in data and choosing a biased human expert for any given input or certain input categories can be actively harmful to the individuals corresponding to those inputs. As such, it is important to ensure that a diverse pool of human experts is chosen to assist with deferred decisions; diversity in the expert pool is desired with respect to both their domains of expertise and their demographics or background. Employing fairness mechanisms can further ensure that the learning algorithm penalizes experts for input categories where they make incorrect decisions due to their biases.

Real-world benchmark dataset. We created an MTurk dataset for offensive language detection to evaluate human-in-the-loop prediction frameworks with multiple experts. The goal of constructing this dataset was to facilitate the learning and evaluation of hybrid frameworks, since having a large number of annotations for each input better enables a learning procedure to differentiate between annotators with different abilities. Existing datasets have often released only aggregate labels, such as by majority voting, which supports ML model training but does not allow modeling individual annotators. To be able to release such data, we have replaced annotator platform IDs with automatically generated pseudonyms.

Our new dataset has important limitations. First, in order to obtain a large number of annotations for each Twitter post, we kept the dataset size relatively small. Furthermore, since the dataset is a subset of the dataset constructed by Davidson et al. [80], it cannot be considered representative of the larger population of Twitter posts/users and the performance demonstrated in our simulations may not translate to larger Twitter datasets. The number of human annotators (170) in our survey is also larger than desired, even though each annotator labels 50-100 posts. Our framework aims to learn the domain of expertise of human experts using only the prior decisions of the experts. However, it is not completely clear how many prior decisions are needed to accurately determine the domain of expertise of every annotator. The gap between the performance using synthetic experts (Section 6.3) and real-world experts (Section 6.4) partially shows that it might be necessary to get more predictions for each expert.

Poursabzi-Sangdeh et al. [248], in a position paper on human-in-the-loop frameworks in facial recognition, argues the necessity of real-world empirical studies of such frameworks to justify their widespread use. They also list the technical challenges associated with such empirical studies. The real-world dataset we provide attempts to initiate a real-world empirical study of human-in-the-loop frameworks

for content moderation but, at the same time, faces similar challenges as highlighted by Poursabzi-Sangdeh et al. [248], i.e., issues with data availability and generalizability of participants/context.

MTurk experiment generalizability. Similar to any other study done using MTurk participants, questions can be raised about the generalizability of the results to a larger population. While MTurk participants do seem suitable for detecting offensive language in Twitter posts (as seen from the performance of the *aggregated crowdworker* in Section 6.4), they may not accurately represent how a lay person would respond to a similar survey or how a domain expert would judge the same posts. The performance of domain experts (people with more experience in screening offensive language) will most likely be better than the accuracy of an average crowd annotator. Correspondingly, our framework with better-trained content moderation experts can be expected to have similar or better performance. Nevertheless, as pointed out in prior work [248, 10], experimental design and choice of participants will play a much bigger role in simulating human-in-the-loop frameworks in settings where human experts cannot be imitated by volunteers.

Addition/removal of experts. An extension of our model that can be further explored is the addition/removal of experts. If a new expert is added to the pipeline and the domain of expertise of this expert is different than the domain of the replaced/existing experts, then the framework might need to be retrained to appropriately include the new expert. This overhead of retraining can, however, be avoided. For instance, one could train the framework using a *basis of experts*, i.e., divide the feature space into interpretable sub-domains and map the experts to these sub-domains. Then if we train the framework using sample decisions of experts with disjoint sub-domains of expertise, we can ensure that the entire feature space is covered either by the classifier or the deferrer (in a similar manner

as Section 6.3.1), and any new expert could be mapped to the corresponding sub-domain. Approaches from prior work [283, 201] can be potentially used to learn these sub-domains and extend our joint learning framework for such settings.

Improved implementation. Like other complex frameworks involving many decision making components, our framework can also suffer from issues that arise from real-world implementations. For instance, dropout reduces overdependence on any particular expert but does not consider the load on any small subset of experts. Alternate load distribution techniques (e.g., Nguyen et al. [228]) can be explored further, at the risk of inducing larger committee sizes. Another extension that can be pursued is to keep the committee size small but variable; this can help with load distribution as well as better committee selection.[2]

[2]The code and dataset for this chapter are available at `https://github.com/vijaykeswani/Deferral-To-Multiple-Experts`.

Chapter 7

Conclusion

The methods proposed in this dissertation provide interventions to incorporate diversity and domain expertise in the outputs of automated decision-making frameworks. For all learning paradigms studied in this dissertation, stakeholder participation consistently improves the performance of the decision-making framework by enhancing the diversity of the output and by using human support in a careful manner to assist automated decision-making.

Chapter 3 forwards an algorithm, *DivScore*, to audit the diversity of any given collection using a small control set (i.e., user-defined representative examples). Theoretical analysis shows that *DivScore* approximates the disparity of the collection, given appropriate control sets and similarity metrics. Empirical evaluations demonstrate that *DivScore* can handle collections from both image and text domains. Crucially, this method allows us to efficiently audit data streams for which protected attribute labels are unavailable.

Chapters 4 and 5 extend the use of representative examples to debias image and text summaries respectively. In both chapters, we first show that current summarization approaches often do not generate summaries that appropriately represent the underlying population distribution. For Google Image Search, we observe how

search results continue to over-represent stereotypical images associated with various occupations. For text summarization, we show that standard summarization algorithms often return summaries that are dialect biased. The approaches presented in these chapters (*QS-balanced* and *MMR-balanced*) aim to ensure fairness in summarization algorithms in the absence of labeled data. As a post-processing approach, our algorithms are also flexible in that they can be applied post-hoc to an existing system where the only additional input necessary is a small set of *diverse* domain-relevant images in the case of image summarization or a small set of *diverse* domain-relevant sentences in the case of text summarization. Due to the generality and simplicity of our approach, these algorithms are expected to perform well for a variety of domains, and it would be interesting to see to what extent they can be applied in areas beyond image and text summarization.

In Chapter 6, we proposed a human-in-the-loop learning model to simultaneously train a classifier and a deferrer in the multiple-experts setting. Theoretical analysis and empirical results for offensive language detection show that this framework, and its fair variants, are able to choose input-specific experts to improve the accuracy and fairness of the decision-making pipeline. This framework can help increase the applicability of automated models in settings where human experts are an indispensable part of the pipeline. Further, using a set of domain experts that is diverse and representative of the underlying population along with fairness mechanisms can ensure that the framework addresses the biases of the model and the humans and that its utilization is thoughtful and context-aware.

The common theme across all chapters is the focus on stakeholder participation. For bias audit and fair summarization, our proposed algorithms utilize user feedback to effectively measure representation disparity and reduce said disparity in automatically-generated summaries. For human-in-the-loop deferral learning, we demonstrate how the heterogeneity of domain experts can be exploited to im-

prove the accuracy and fairness of human-assisted decision-making systems. In all settings, stakeholder involvement (either as users or domain experts) provides additional information that improves the framework's performance. Importantly, this additional information is often unavailable to or under-utilized by the framework through the data it is trained on. As such, stakeholder involvement in automated frameworks adds an additional dimension along which we can incorporate human decision-making values that are absent from the available data.

At the beginning of Chapter 1, I talked about how the rationality of a decision-making process is dependent upon the values of the decision-maker. In the case of an automated decision-making framework, we cannot point to one person and say that the decisions reflect their values. Instead, there are multiple human stakeholders involved throughout the process of designing, developing, assisting, and deploying an automated decision-making framework. The values reflected in such frameworks are derived from all of these stakeholders as well as from the institutional values of the parent organization. As such, when we talk about the problem of social biases in automated decisions, we are not just pointing to the prejudices of certain human decision-makers, but also the structural prejudices of the developing institution and those encoded within the data and model used by the automated decision-making framework. The presence of such biases points to a mismatch between the values of the framework developers/institutions and the values of users of the framework, and, correspondingly, leads to reduced performance (in the form of misrepresentation or disparate impact) for users from systematically-disadvantaged demographic groups. Hence, it will always be beneficial to encourage users to participate by providing feedback or assisting the decision-making framework, in a manner that allows them to share their values with the framework. Eliciting diverse voices during development and deployment can allow us to understand and incorporate common ethical principles in

automated decision-making frameworks and take steps toward building trust in these frameworks. That is indeed the goal of the methods proposed in this dissertation.

As the final point, I believe it is important to mention that there are many other dimensions of decision-making along which stakeholder participation can be useful. Users can also be consulted during the design of decision-making frameworks and using aggregation methods from social choice theory can allow the identification of the framework components important to different groups of users [273, 110, 300]. Participatory action research similarly emphasizes the collective development of decision-making frameworks where the experience and knowledge of diverse stakeholders are explicitly solicited during the design of socially-relevant systems [114, 137, 184, 176]. Abiding by the principles highlighted in these fields of research can significantly improve the performance of automated decision-making frameworks and potentially alleviate many concerns regarding the impact of these frameworks.

Through this dissertation, I have highlighted crucial areas where algorithmic development falls short of creating progressive frameworks and suggested mechanisms by which we can modify such frameworks to obtain unbiased and accurate outcomes through stakeholder involvement. Implementing these frameworks in real-world applications will face many more challenges; nevertheless, taking a participatory approach to address these challenges can help ensure that the impact of automation on our society is equitable.

Bibliography

[1] Risk, Race, & Recidivism: Predictive Bias and Disparate Impact.(2016).

[2] Bureau of Labor Statistics. Labor Force Statistics from the Current Population Survey. https://www.bls.gov/cps/aa2012/cpsaat11.htm, 2013.

[3] When It Comes to Gorillas, Google Photos Remains Blind. https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/, 2018.

[4] IBM Response to "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". http://gendershades.org/docs/ibm.pdf, 2018.

[5] Gender and Jobs in Online Image Searches. https://www.pewsocialtrends.org/2018/12/17/gender-and-jobs-in-online-image-searches/, 2018.

[6] Appel Citoyen. https://appelcitoyen.ch/on-ouvre-les-urnes-donnees-brutes-de-la-primaire/, 2018.

[7] AI reveals misrepresentation of engineers online. https://www.raeng.org.uk/news/news-releases/2019/november/ai-reveals-misrepresentation-of-engineers-online, 2019.

[8] The Secret Bias Hidden in Mortgage-Approval Algorithms. https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms, 2021.

[9] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, pages 60–69, 2018.

[10] Alfredo Alba, Anni Coden, Anna Lisa Gentile, Daniel Gruhl, Petar Ristoski, and Steve Welch. Multi-lingual concept extraction with linked data and human-in-the-loop. In *Proceedings of the Knowledge Capture Conference*, pages 1–8, 2017.

[11] Eugenio Alberdi, Lorenzo Strigini, Andrey A Povyakalo, and Peter Ayton. Why are people's decisions sometimes worse with computer support? In *International Conference on Computer Safety, Reliability, and Security*, pages 18–31. Springer, 2009.

[12] Rasim M Alguliev, Ramiz M Aliguliyev, Makrufa S Hajirahimova, and Chingiz A Mehdiyev. MCMR: Maximum coverage and minimum redundant text summarization model. *Expert Systems with Applications*, 38(12), 2011.

[13] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica*, 2016.

[14] Luca Anzalone, Paola Barra, Silvio Barra, Fabio Narducci, and Michele Nappi. Transfer Learning for Facial Attributes Prediction and Clustering. In *International Conference on Smart City and Informatization*, pages 105–117. Springer, 2019.

[15] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR 2017*, 2017.

[16] Pranjal Awasthi, Matthaus Kleindessner, and Jamie Morgenstern. Equalized odds postprocessing under imperfect group information. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020*, 2020.

[17] Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. Scalable fair clustering. In *International Conference on Machine Learning*, pages 405–413. PMLR, 2019.

[18] Eric Bair. Semi-supervised clustering methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(5):349–361, 2013.

[19] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. Optimizing AI for Teamwork. *arXiv preprint arXiv:2004.13102*, 2020.

[20] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California Law Review*, 2016.

[21] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2, 2017.

[22] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. URL `http://www.fairmlbook.org`.

[23] Julia B Bear, Lily Cushenbery, Manuel London, and Gary D Sherman. Performance feedback, power retention, and the gender gap in leadership. *The Leadership Quarterly*, 28(6):721–740, 2017.

[24] Gary S Becker. *The economics of discrimination*. University of Chicago press, 2010.

[25] Kyla Bender-Baird. Peeing under surveillance: bathrooms, gender policing, and hate violence. *Gender, Place & Culture*, 23(7):983–988, 2016.

[26] Cynthia L Bennett and Os Keyes. What is the Point of Fairness? Disability, AI and The Complexity of Justice. In *ASSETS 2019 Workshop—AI Fairness for People with Disabilities*, 2019.

[27] Camiel J Beukeboom and Christian Burgers. How stereotypes are shared through language: a review and introduction of the aocial categories and stereotypes communication framework. *Review of Communication Research*, 2019.

[28] Su Lin Blodgett and Brendan O'Connor. Racial disparity in natural language processing: A case study of social media african-american english. 2017.

[29] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2016.

[30] Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. A dataset and classifier for recognizing social media english. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 56–61, 2017.

[31] Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. Twitter universal dependency parsing for African-American and mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, 2018.

[32] Su Lin Blodgett, Solon Barocas, Hal Daumé, III, and Hanna Wallach. Language (technology) is Power: A Critical Survey of "Bias" in NLP. In *Pro-*

ceedings of the Conference of the Association for Computational Linguistics (ACL), 2020.

[33] Galen V Bodenhausen and Robert S Wyer. Effects of stereotypes in decision making and information-processing strategies. *Journal of personality and social psychology*, 48(2):267, 1985.

[34] 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. ISSN 2307-387X.

[35] 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.

[36] Amanda Bower, Sarah N Kitchen, Laura Niss, Martin J Strauss, Alexander Vargas, and Suresh Venkatasubramanian. Fair pipelines. *arXiv preprint arXiv:1707.00391*, 2017.

[37] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[38] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71:102062, 2021.

[39] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[40] Robin Burke, Alexander Felfernig, and Mehmet H Göker. Recommender systems: An overview. *Ai Magazine*, 32(3):13–18, 2011.

[41] Mara Cadinu, Marcella Latrofa, and Andrea Carnaghi. Comparing Self-stereotyping with In-group-stereotyping and Out-group-stereotyping in Unequal-status Groups: The Case of Gender. *Self and Identity*, 12(6):582–596, 2013.

[42] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building Classifiers with Independency Constraints. *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18, 2009.

[43] Gruia Calinescu, Chandra Chekuri, Martin Pál, and Jan Vondrák. Maximizing a submodular set function subject to a matroid constraint. In *International Conference on Integer Programming and Combinatorial Optimization*, pages 182–196. Springer, 2007.

[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] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

[46] Jaime G Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, volume 98, pages 335–336, 1998.

[47] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of

supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168, 2006.

[48] M Emre Celebi and Kemal Aydin. *Unsupervised learning algorithms*, volume 9. Springer, 2016.

[49] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of Text Generation: A Survey. *arXiv preprint arXiv:2006.14799*, 2020.

[50] L Elisa Celis and Vijay Keswani. Implicit Diversity in Image Summarization. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–28, 2020.

[51] L. Elisa Celis, Amit Deshpande, Tarun Kathuria, and Nisheeth K Vishnoi. How to be fair and diverse? *arXiv preprint arXiv:1610.07183*, 2016.

[52] L. Elisa Celis, Vijay Keswani, Damian Straszak, Amit Deshpande, Tarun Kathuria, and Nisheeth Vishnoi. Fair and Diverse DPP-Based Data Summarization. In *International Conference on Machine Learning*, pages 715–724, 2018.

[53] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. Ranking with Fairness Constraints. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[54] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 319–328. ACM, 2019.

[55] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi.

Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *FAT* 2019*, pages 319–328, 2019.

[56] L Elisa Celis, Vijay Keswani, and Nisheeth Vishnoi. Data preprocessing to mitigate bias: A maximum entropy based approach. In *International Conference on Machine Learning*, 2020.

[57] Stevie Chancellor, Shion Guha, Jofish Kaye, Jen King, Niloufar Salehi, Sarita Schoenebeck, and Elizabeth Stowell. The Relationships between Data, Power, and Justice in CSCW Research. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, pages 102–105, 2019.

[58] Abdelhamid Chellal and Mohand Boughanem. Optimization framework model for retrospective tweet summarization. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pages 704–711, 2018.

[59] Lingjiao Chen, Matei Zaharia, and James Zou. FrugalML: How to use ML Prediction APIs more accurately and cheaply. In *NeurIPS*, 2020.

[60] Xingyu Chen, Brandon Fain, Liang Lyu, and Kamesh Munagala. Proportionally fair clustering. In *International Conference on Machine Learning*, pages 1032–1041. PMLR, 2019.

[61] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. *Advances in neural information processing systems*, 30, 2017.

[62] Kristy Choi, Aditya Grover, Rui Shu, and Stefano Ermon. Fair Generative Modeling via Weak Supervision. In *ICML*, 2020.

[63] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *ICML*. PMLR, 2020.

[64] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[65] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148, 2018.

[66] Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666. ACM, 2008.

[67] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[68] Patricia Hill Collins. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. routledge, 2002.

[69] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[70] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

[71] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. *Advances in Neural Information Processing Systems*, 29:1660–1668, 2016.

[72] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pages 67–82. Springer, 2016.

[73] Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Scott Yang. Online learning with abstention. In *International conference on machine learning*, pages 1059–1067. PMLR, 2018.

[74] Donna Crawley. Gender and perceptions of occupational prestige: Changes over 20 years. *Sage Open*, 4(1):2158244013518923, 2014.

[75] Mary Cummings. Automation bias in intelligent time critical decision support systems. In *AIAA 1st Intelligent Systems Technical Conference*, page 6313, 2004.

[76] Antitza Dantcheva and François Brémond. Gender estimation based on smile-dynamics. *IEEE Transactions on Information Forensics and Security*, 12 (3):719–729, 2016.

[77] Abhisek Dash, Anurag Shandilya, Arindam Biswas, Kripabandhu Ghosh, Saptarshi Ghosh, and Abhijnan Chakraborty. Summarizing user-generated textual content: Motivation and methods for fairness in algorithmic summaries. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–28, 2019.

[78] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-`

that-showed-bias-against-women-idUSKCN1MK08G, 2018.

[79] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

[80] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*, 2017.

[81] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, 2019.

[82] Peter Dayan, Maneesh Sahani, and Grégoire Deback. Unsupervised learning. *The MIT encyclopedia of the cognitive sciences*, pages 857–859, 1999.

[83] Abir De, Paramita Koley, Niloy Ganguly, and Manuel Gomez-Rodriguez. Regression under human assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2611–2620, 2020.

[84] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[85] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[86] Mark DeSantis and Nathan Sierra. Women smiled more often and openly

than men when photographed for a pleasant, public occasion in 20 (th) century United States society. *Psychology*, 37(3-4):21–31, 2000.

[87] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[88] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax Group Fairness: Algorithms and Experiments. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2021.

[89] William Dieterich, Christina Mendoza, and Tim Brennan. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc*, 2016.

[90] Carl DiSalvo, Andrew Clement, and Volkmar Pipek. Communities: Participatory Design for, with and by communities. In *Routledge international handbook of participatory design*, pages 202–230. Routledge, 2012.

[91] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

[92] John F Dovidio, Susan Eggly, Terrance L Albrecht, Nao Hagiwara, and Louis A Penner. Racial biases in medicine and healthcare disparities. *TPM: Testing, Psychometrics, Methodology in Applied Psychology*, 23(4), 2016.

[93] Gabriel Doyle. Mapping dialectal variation by querying social media. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 98–106, 2014.

[94] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

[95] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[96] Grant Duwe and Michael Rocque. Effects of Automating Recidivism Risk Assessment on Reliability, Predictive Validity, and Return on Investment (ROI). *Criminology & Public Policy*, 16(1):235–269, 2017.

[97] Cynthia Dwork and Christina Ilvento. Fairness Under Composition. In *Innovations in Theoretical Computer Science Conference (ITCS)*, 2019.

[98] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[99] Lilian Edwards and Michael Veale. Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.*, 16:18, 2017.

[100] Max Ehrlich, Timothy J Shields, Timur Almaev, and Mohamed R Amer. Facial attributes classification using multi-task representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 47–55, 2016.

[101] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1277–1287, 2010.

[102] Ran El-Yaniv et al. On the Foundations of Noise-free Selective Classification. *Journal of Machine Learning Research*, 11(5), 2010.

[103] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*, pages 160–171. PMLR, 2018.

[104] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 2004.

[105] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[106] Evanthia Faliagka, Kostas Ramantas, Athanasios Tsakalidis, and Giannis Tzimas. Application of machine learning algorithms to an online recruitment system. In *Proc. International Conference on Internet and Web Applications and Services*, pages 215–220, 2012.

[107] Hanming Fang and Andrea Moro. Theories of statistical discrimination and affirmative action: A survey. *Handbook of social economics*, 1:133–200, 2011.

[108] Moran Feldman, Amin Karbasi, and Ehsan Kazemi. Do less, get more: streaming submodular maximization with subsampling. In *Advances in Neural Information Processing Systems*, pages 732–742, 2018.

[109] Eimear Finnegan, Jane Oakhill, and Alan Garnham. Counter-stereotypical pictures as a strategy for overcoming spontaneous gender stereotypes. *Frontiers in psychology*, 6:1291, 2015.

[110] Jessie Finocchiaro, Roland Maio, Faidra Monachou, Gourab K Patro, Manish Raghavan, Ana-Andreea Stoica, and Stratis Tsirtsis. Bridging machine learning and mechanism design towards algorithmic fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 489–503, 2021.

[111] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining, 2016*, pages 144–152. SIAM, 2016.

[112] T Fitzpatrick. Fitzpatrick Skin Type Classification Scale. *Skin Inc*, 2008.

[113] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 329–338, 2019.

[114] Batya Friedman. Value-sensitive design. *interactions*, 3(6):16–23, 1996.

[115] Batya Friedman, Peter H Kahn, and Alan Borning. Value sensitive design and information systems. *The handbook of information and computer ethics*, pages 69–101, 2008.

[116] Tamar Szabó Gendler. On the epistemic costs of implicit bias. *Philosophical Studies*, 156:33–63, 2011.

[117] George Gerbner, Larry Gross, Michael Morgan, and Nancy Signorielli. Living with television: The dynamics of the cultivation process. *Perspectives on media effects*, 1986:17–40, 1986.

[118] Patricia Gherovici. *Please select your gender: From the invention of hysteria to the*

democratizing of transgenderism. Routledge, 2011.

[119] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, 2012.

[120] Fréderic Godin. *Improving and Interpreting Neural Networks for Word-Level Prediction Tasks in Natural Language Processing*. PhD thesis, Ghent University, Belgium, 2019.

[121] Fréderic Godin. Improving and Interpreting Neural Networks for Word-Level Prediction Tasks in Natural Language Processing. *Ghent University*, 2019.

[122] Jade Goldstein and Jaime Carbonell. Summarization: Using MMR for Diversity-Based Reranking and Evaluating Summaries. Technical report, Carnegie-Mellon Univ PA Language Technology Inst, 1998.

[123] Stephen Gorard. Revisiting a 90-year-old debate: The Advantages of the Mean Deviation. *British Journal of Educational Studies*, 53(4):417–430, 2005.

[124] Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Loubes Jean-Michel. Obtaining Fairness using Optimal Transport Theory. In *International Conference on Machine Learning*, pages 2357–2365, 2019.

[125] Nitesh Goyal, Ian D Kivlichan, Rachel Rosen, and Lucy Vasserman. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–28, 2022.

[126] Ben Green. "Good" isn't good enough. In *Proceedings of the AI for Social Good*

workshop at NeurIPS, 2019.

[127] Ben Green and Yiling Chen. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 90–99, 2019.

[128] Nina Grgić-Hlača, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018*, pages 903–912, 2018.

[129] Tor Grønsund and Margunn Aanestad. Augmenting the Algorithm: Emerging human-in-the-loop work configurations. *The Journal of Strategic Information Systems*, 29(2):101614, 2020.

[130] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. Who Said What: Modeling Individual Labelers Improves Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[131] Kilem L Gwet. On The Krippendorff's Alpha Coefficient. 2011. URL `https://agreestat.com/papers/onkrippendorffalpha_rev10052015.pdf`.

[132] James Hafner, Harpreet S. Sawhney, William Equitz, Myron Flickner, and Wayne Niblack. Efficient color histogram indexing for quadratic form distance functions. *IEEE transactions on pattern analysis and machine intelligence*, 17(7):729–736, 1995.

[133] Aaron Halfaker and R Stuart Geiger. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. *arXiv preprint arXiv:1909.05189*, 2019.

[134] Jens Hälterlein. Epistemologies of predictive policing: Mathematical social science, social physics and machine learning. *Big data & society*, 8(1): 20539517211003118, 2021.

[135] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 501–512, 2020.

[136] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[137] Christina Harrington, Sheena Erete, and Anne Marie Piper. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25, 2019.

[138] Trudier Harris. *From mammies to militants: Domestics in black American literature*. Temple University Press, 1982.

[139] S Alexander Haslam, John C Turner, Penelope J Oakes, Katherine J Reynolds, and Bertjan Doosje. From personal pictures in the head to collective tools in the world: How shared stereotypes allow groups to represent and change social reality. 2002.

[140] Trevor Hastie, Robert Tibshirani, Jerome Friedman, Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Overview of supervised learning. *The elements of statistical learning: Data mining, inference, and prediction*, pages 9–41, 2009.

[141] Ruifang He and Xingyi Duan. Twitter summarization based on social network and sparse reconstruction. In *Thirty-Second AAAI Conference on AI*, 2018.

[142] Madeline E Heilman, Francesca Manzi, and Susanne Braun. Presumed incompetent: Perceived lack of fit and gender bias in recruitment and selection. In *Handbook of gendered careers in management*, pages 90–104. Edward Elgar Publishing, 2015.

[143] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 793–811. Springer, 2018.

[144] Jody L Herman. Gendered restrooms and minority stress: The public regulation of gender and its impact on transgender people's lives. *Journal of Public Management & Social Policy*, 19(1):65, 2013.

[145] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701, 2015.

[146] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms I: Fundamentals*, volume 305. Springer science & business media, 2013.

[147] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[148] Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. Understand-

ing US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59:244–255, 2016.

[149] Mara Hvistendahl. Can "predictive policing" prevent crime before it happens. *Science Magazine*, 28, 2016.

[150] David Inouye and Jugal K Kalita. Comparing twitter summarization algorithms for multiple post summaries. In *2011 IEEE Third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 298–306. IEEE, 2011.

[151] Aishwarya Jadhav and Vaibhav Rajan. Extractive summarization with swapnet: Sentences and words from alternating pointer networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

[152] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4), 2017.

[153] Taylor Jones, Jessica Rose Kalbfeld, Ryan Hancock, and Robin Clark. Testifying while black: An experimental study of court reporter accuracy in transcription of African American English. *Language*, 95(2):e216–e252, 2019.

[154] Anna Jørgensen, Dirk Hovy, and Anders Søgaard. Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18, 2015.

[155] Stephanie Julia Kapusta. Misgendering and its moral contestability. *Hypatia*, 31(3):502–519, 2016.

[156] Hyun Joon Jung and Matthew Lease. Crowdsourced Task Routing via Ma-

trix Factorization. *arXiv preprint arXiv:1310.5142*, 2013.

[157] Hyun Joon Jung, Yubin Park, and Matthew Lease. Predicting Next Label Quality: A Time-Series Model of Crowdwork. *HCOMP*, 14:1–9, 2014.

[158] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 110–110, 2020.

[159] Ece Kamar, Ashish Kapoor, and Eric Horvitz. Identifying and Accounting for Task-dependent Bias in Crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 3, 2015.

[160] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*, pages 1–6. IEEE, 2009.

[161] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1): 1–33, 2012.

[162] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.

[163] Dana Kanze, Laura Huang, Mark A Conley, and E Tory Higgins. We ask men to win and women not to lose: Closing the gender gap in startup funding. *Academy of Management Journal*, 61(2):586–614, 2018.

[164] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architec-

ture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[165] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and PM Krafft. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 45–55, 2020.

[166] Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3819–3828. ACM, 2015.

[167] Vijay Keswani and L Elisa Celis. Dialect Diversity in Text Summarization on Twitter. In *Proceedings of the Web Conference 2021*, pages 3802–3814, 2021.

[168] Vijay Keswani and L Elisa Celis. Auditing for Diversity using Representative Examples. *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2021.

[169] Vijay Keswani and L Elisa Celis. An Anti-Subordination Approach to Fair Classification. 2022.

[170] Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Towards Unbiased and Accurate Deferral to Multiple Experts. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2021.

[171] Peter Kieseberg, Edgar Weippl, and Andreas Holzinger. Trust for the doctor-in-the-loop. *ERCIM news*, 104(1):32–33, 2016.

[172] Svetlana Kiritchenko and Saif M Mohammad. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. *NAACL HLT 2018*, 2018.

[173] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, pages 43:1–43:23, 2017.

[174] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015.

[175] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. Neural Text Summarization: A Critical Evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019.

[176] Sarah Kuhn and Michael J Muller. Participatory design. *Communications of the ACM*, 36(6):24–29, 1993.

[177] Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

[178] Sabine Landau, Morven Leese, Daniel Stahl, and Brian S Everitt. *Cluster Analysis*. John Wiley & Sons, 2011.

[179] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[180] Amy N Langville and Carl D Meyer. Google's PageRank and beyond. In

Google's PageRank and Beyond. Princeton university press, 2011.

[181] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

[182] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)*, 9, 2016.

[183] Makeba Lavan. The Negro Tweets His Presence: Black Twitter as Social and Political Watchdog. *Modern Language Studies*, pages 56–65, 2015.

[184] Christopher A Le Dantec and Sarah Fox. Strangers at the gate: Gaining access, building rapport, and co-constructing community-based research. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 1348–1358, 2015.

[185] Christopher A Le Dantec, Erika Shehan Poole, and Susan P Wyche. Values as lived experience: evolving value sensitive design in support of value discovery. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1141–1150, 2009.

[186] Matthew Lease. On quality control and machine learning in crowdsourcing. *Human Computation*, 11(11), 2011.

[187] Ju-Hong Lee, Sun Park, Chan-Min Ahn, and Daeho Kim. Automatic generic document summarization based on non-negative matrix factorization. *Information Processing & Management*, 45(1):20–34, 2009.

[188] Gil Levi and Tal Hassner. Emotion recognition in the wild via convolu-

tional neural networks and mapped binary patterns. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 503–510. ACM, 2015.

[189] Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081, 2018.

[190] Hongwei Li and Qiang Liu. Cheaper and Better: Selecting Good Workers for Crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 3, 2015.

[191] Lihong Li, Michael L Littman, Thomas J Walsh, and Alexander L Strehl. Knows what it knows: a framework for self-aware learning. *Machine learning*, 82(3):399–443, 2011.

[192] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157, 2003.

[193] Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 510–520. Association for Computational Linguistics, 2011.

[194] Hui Lin and Vincent Ng. Abstractive Summarization: A Survey of the State of the Art. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[195] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt.

Delayed impact of fair machine learning. In *ICML*, pages 3150–3158, 2018.

[196] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6196–6200. AAAI Press, 2019.

[197] Yang Liu and Mirella Lapata. Text Summarization with Pretrained Encoders. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, 2019.

[198] Zimo Liu, Jingya Wang, Shaogang Gong, Huchuan Lu, and Dacheng Tao. Deep reinforcement active learning for human-in-the-loop person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6122–6131, 2019.

[199] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[200] Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Deep gamblers: Learning to abstain with portfolio theory. In *Advances in Neural Information Processing Systems*, pages 10623–10633, 2019.

[201] Pedro Lopez-Garcia, Antonio D Masegosa, Eneko Osaba, Enrique Onieva, and Asier Perallos. Ensemble classification for imbalanced data based on feature space partitioning and hybrid metaheuristics. *Applied Intelligence*, 49 (8):2807–2822, 2019.

[202] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam

Datta. Gender bias in neural natural language processing. *arXiv preprint arXiv:1807.11714*, 2018.

[203] Hans Peter Luhn. A statistical approach to mechanized encoding and searching of literary information. In *IBM Journal of research and development*, 1957.

[204] David Madras, Toni Pitassi, and Richard Zemel. Predict Responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, pages 6147–6157, 2018.

[205] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax Pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pages 6755–6764. PMLR, 2020.

[206] Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[207] Craig McGarty, Vincent Y Yzerbyt, Russel Spears, et al. Social, cultural and cognitive factors in stereotype formation. *Stereotypes as explanations: The formation of meaningful beliefs about social groups*, 1:1–16, 2002.

[208] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[209] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the conference on empirical methods in natural language processing*, 2004.

[210] Tomas Mikolov, Kai Chen, Gregory S Corrado, and Jeffrey A Dean. Computing numeric representations of words in a high-dimensional space, May 19 2015. US Patent 9,037,464.

[211] Claire Cain Miller. Can an algorithm hire better than a human. *The New York Times*, 25, 2015.

[212] Derek Miller. Leveraging BERT for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*, 2019.

[213] Zachary Miller, Brian Dickinson, and Wei Hu. Gender prediction on twitter using stream algorithms with n-gram character features. 2012.

[214] Baharan Mirzasoleiman, Stefanie Jegelka, and Andreas Krause. Streaming non-monotone submodular maximization: Personalized video summarization on the fly. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[215] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

[216] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT press, 2018.

[217] Ellis P Monk. The color of punishment: African Americans, skin tone, and the criminal justice system. *Ethnic and Racial Studies*, 42(10):1593–1612, 2019.

[218] Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pages 7076–7087. PMLR, 2020.

[219] Hussein Mozannar, Arvind Satyanarayan, and David Sontag. Teaching hu-

mans when to defer to a classifier via exemplars. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5323–5331, 2022.

[220] Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, 2019.

[221] Michael Muller, Cecilia Aragon, Shion Guha, Marina Kogan, Gina Neff, Cathrine Seidelin, Katie Shilton, and Anissa Tanweer. Interrogating Data Science. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, pages 467–473, 2020.

[222] Michael J Muller. Participatory design: the third space in HCI. In *The human-computer interaction handbook*, pages 1087–1108. CRC press, 2007.

[223] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models, 2020.

[224] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[225] Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, volume 1170, 2018.

[226] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*, 2018.

[227] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical programming*, 14(1):265–294, 1978.

[228] An Thanh Nguyen, Byron C Wallace, and Matthew Lease. Combining crowd and expert labels using decision theoretic active learning. In *Third AAAI conference on human computation and crowdsourcing*, 2015.

[229] Hieu V Nguyen and Li Bai. Cosine similarity metric learning for face verification. In *Asian conference on computer vision*, pages 709–720. Springer, 2010.

[230] Minh-Tien Nguyen, Dac Viet Lai, Huy Tien Nguyen, and Minh Le Nguyen. Tsix: a human-involved-creation dataset for tweet summarization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018.

[231] Beibei Niu, Jinzheng Ren, and Xiaotao Li. Credit scoring using machine learning by combing social network information: Evidence from peer-to-peer lending. *Information*, 10(12):397, 2019.

[232] Safiya U. Noble. *Algorithms of oppression: How search engines reinforce racism*. NYU Press, 2018.

[233] Northpointe. Compas risk and need assessment systems. `http://www.northpointeinc.com/files/downloads/FAQ_Document.pdf`, 2012.

[234] Besmira Nushi, Adish Singla, Anja Gruenheid, Erfan Zamanian, Andreas Krause, and Donald Kossmann. Crowd Access Path Optimization: Diversity Matters. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 3, 2015.

[235] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 2019.

[236] Alexandra Olteanu, Kartik Talamadupula, and Kush R Varshney. The limits of abstract evaluation metrics: The case of hate speech detection. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 405–406, 2017.

[237] Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[238] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014.

[239] Margaret Ott. Tweet like a girl: A corpus analysis of gendered language in social media. *Yale University, apr*, 2016.

[240] Makbule Gulcin Ozsoy, Ferda Nur Alpaslan, and Ilyas Cicekli. Text summarization using latent semantic analysis. *Journal of Information Science*, 2011.

[241] Aishwarya Padmakumar and Akanksha Saran. Unsupervised Text Summarization Using Sentence Embeddings. Technical report, Technical Report, University of Texas at Austin, 2016.

[242] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of EMNLP 2018*, 2018.

[243] Genevieve Patterson, Grant Van Horn, Serge Belongie, Pietro Perona, and James Hays. Bootstrapping Fine-Grained Classifiers: Active Learning with

a Crowd in the Loop. In *NeurIPS Workshop on Crowdsourcing: Theory, Algorithms and Applications*, 2013.

[244] 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.

[245] Claudia Perlich, Brian Dalessandro, Troy Raeder, Ori Stitelman, and Foster Provost. Machine learning for targeted display advertising: Transfer learning in action. *Machine learning*, 95(1):103–127, 2014.

[246] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

[247] W James Potter. Cultivation theory and research: A methodological critique. *Journalism & Mass Communication Monographs*, (147):1, 1994.

[248] Forough Poursabzi-Sangdeh, Samira Samadi, Jennifer Wortman Vaughan, and Hanna Wallach. A Human in the Loop is Not Enough: The Need for Human-Subject Experiments in Facial Recognition. In *CHI Workshop on Human-Centered Approaches to Fair and Responsible AI*, 2020.

[249] Chenxi Qiu, Anna C Squicciarini, Barbara Carminati, James Caverlee, and Dev Rishi Khare. Crowdselect: Increasing Accuracy of Crowdsourcing Tasks through Behavior Prediction and User Selection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 539–548, 2016.

[250] Dragomir R Radev, Sasha Blair-Goldensohn, and Zhu Zhang. Experiments

in single and multidocument summarization using MEAD. In *First document understanding conference*, page 1À8. Citeseer, 2001.

[251] Filip Radlinski, Paul N Bennett, Ben Carterette, and Thorsten Joachims. Redundancy, diversity and interdependent document relevance. In *ACM SIGIR Forum*, volume 43, pages 46–52. ACM, 2009.

[252] Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*, 2019.

[253] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Bobby Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. Direct uncertainty prediction for medical second opinions. In *International Conference on Machine Learning*, pages 5281–5290, 2019.

[254] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 145–151, 2020.

[255] Tiziana Ramaci, Monica Pellerone, Caterina Ledda, Giovambattista Presti, Valeria Squatrito, and Venerando Rapisarda. Gender stereotypes in occupational choice: a cross-sectional study on a group of Italian adolescents. *Psychology Research and Behavior Management*, 10:109, 2017.

[256] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9): 2352–2449, 2017.

[257] Kan Ren, Weinan Zhang, Ke Chang, Yifei Rong, Yong Yu, and Jun Wang.

Bidding machine: Learning to bid for directly optimizing profits in display advertising. *IEEE Transactions on Knowledge and Data Engineering*, 30(4):645–659, 2017.

[258] Willy E Rice. Race, Gender, Redlining, and the Discriminatory Access to Loans, Credit, and Insurance: An Historical and Empirical Analysis of Consumers Who Sued Lenders and Insurers in Federal and State Courts, 1950-1995. *San Diego L. Rev.*, 33:583, 1996.

[259] Rashida Richardson, Jason M Schultz, and Kate Crawford. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev.*, 2019.

[260] John R Rickford. *Raciolinguistics: How language shapes our ideas about race*. Oxford University Press, 2016.

[261] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420, 2007.

[262] Gaetano Rossiello, Pierpaolo Basile, and Giovanni Semeraro. Centroid-based text summarization through compositionality of word embeddings. In *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, pages 12–21, 2017.

[263] Andras Rozsa, Manuel Günther, Ethan M Rudd, and Terrance E Boult. Facial attributes: Accuracy and adversarial robustness. *Pattern Recognition Letters*, 124:100–108, 2019.

[264] 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. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 458–468, 2020.

[265] Elizabeth B-N Sanders. From user-centered to participatory design approaches. In *Design and the social sciences*, pages 18–25. CRC Press, 2002.

[266] Mark Sanderson, Jiayu Tang, Thomas Arni, and Paul Clough. What else is there? search diversity examined. In *European Conference on Information Retrieval*, pages 562–569. Springer, 2009.

[267] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, 2019.

[268] Hannah Sassaman, Jennifer Lee, Jenessa Irvine, and Shankar Narayan. Creating community-based tech policy: case studies, lessons learned, and what technologists and communities can do together. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 685–685, 2020.

[269] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–33, 2019.

[270] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[271] Carsten Schwemmer, Carly Knight, Emily D Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W Lockhart. Diagnosing gender bias in image recognition systems. *Socius*, 6:2378023120967171, 2020.

[272] Andrew D Selbst. Disparate impact in big data policing. *Ga. L. Rev.*, 52:109, 2017.

[273] Amartya Sen. Social choice theory. *Handbook of mathematical economics*, 3: 1073–1181, 1986.

[274] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

[275] Larry J Shrum. Assessing the social influence of television: A social cognition perspective on cultivation effects. *Communication Research*, 22(4):402–429, 1995.

[276] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[277] Raven Sinclair. The Indigenous child removal system in Canada: An examination of legal decision-making and racial bias. *First Peoples Child & Family Review: An Interdisciplinary Journal Honouring the Voices, Perspectives, and Knowledges of First Peoples through Research, Critical Analyses, Stories, Standpoints and Media Reviews*, 11(2):8–18, 2016.

[278] Vivek K Singh, Mary Chayko, Raj Inamdar, and Diana Floegel. Female Librarians and Male Computer Programmers? Gender Bias in Occupational

Images on Digital Media Platforms. *Journal of the Association for Information Science and Technology*, 2020.

[279] Pinaki Sinha and Ramesh Jain. Extractive summarization of personal photos from life events. In *2011 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2011.

[280] Mark Snyder, Elizabeth Decker Tanke, and Ellen Berscheid. Social perception and interpersonal behavior: On the self-fulfilling nature of social stereotypes. *Journal of Personality and social Psychology*, 35(9):656, 1977.

[281] Steven J Spencer, Claude M Steele, and Diane M Quinn. Stereotype threat and women's math performance. *Journal of experimental social psychology*, 35 (1):4–28, 1999.

[282] Eliza Strickland, 2018. URL `https://spectrum.ieee.org/computing/software/aihuman-partnerships-tackle-fake-news`.

[283] Carolin Strobl, James Malley, and Gerhard Tutz. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, 14(4): 323, 2009.

[284] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

[285] Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization*, pages 1–9. 2021.

[286] Andrew Sutton, Reza Samavi, Thomas E Doyle, and David Koff. Digitized trust in human-in-the-loop health research. In *2018 16th Annual Conference on Privacy, Security and Trust (PST)*, pages 1–10. IEEE, 2018.

[287] Henri Tajfel. Social stereotypes and social groups. 2001.

[288] Yi Chern Tan and L Elisa Celis. Assessing social and intersectional biases in contextualized word representations. In *Neural Information Processing Systems*, 2019.

[289] Rachael Tatman. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 2017.

[290] LII Wex Definitions Team. Protected Characteristics. `https://www.law.cornell.edu/wex/protected_characteristic`, 2020.

[291] J Michael Terry, Randall Hendrick, Evangelos Evangelou, and Richard L Smith. Variable dialect switching among African American children: Inferences about working memory. *Lingua*, 120(10):2463–2475, 2010.

[292] Nenad Tomasev, Kevin R McKee, Jackie Kay, and Shakir Mohamed. Fairness for unobserved characteristics: Insights from technological impacts on queer communities. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 254–265, 2021.

[293] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[294] Sebastian Tschiatschek, Rishabh K Iyer, Haochen Wei, and Jeff A Bilmes. Learning mixtures of submodular functions for image collection summariza-

tion. In *Advances in neural information processing systems*, pages 1413–1421, 2014.

[295] Jinzheng Tu, Guoxian Yu, Carlotta Domeniconi, Jun Wang, Guoqiang Xiao, and Maozu Guo. Multi-label Crowd Consensus via Joint Matrix Factorization. *Knowledge and Information Systems*, 62(4):1341–1369, 2020.

[296] Emiel Van Miltenburg. Stereotyping and bias in the flickr30k dataset. *arXiv preprint arXiv:1605.06083*, 2016.

[297] Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 43(6):1606–1618, 2007.

[298] Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. Community-based Bayesian Aggregation Models for Crowdsourcing. In *Proceedings of the 23rd international conference on World wide web*, pages 155–164, 2014.

[299] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.

[300] Salomé Viljoen, Jake Goldenfein, and Lee McGuigan. Design choices: Mechanism design and platform capitalism. *Big data & society*, 8(2): 20539517211034312, 2021.

[301] Hao Wang, Berk Ustun, and Flavio P Calmon. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. *arXiv preprint arXiv:1901.10501*, pages 6618–6627, 2019.

[302] Joel S Weissman and Romana Hasnain-Wynia. Advancing health care equity through improved data collection. *The New England journal of medicine*, 364 (24):2276–2277, 2011.

[303] Kelly Lais Wiggers, Alceu de Souza Britto Junior, Alessandro Lameiras Koerich, Laurent Heutte, and Luiz Eduardo Soares de Oliveira. Deep Learning Approaches for Image Retrieval and Pattern Spotting in Ancient Documents. *arXiv preprint arXiv:1907.09404*, 2019.

[304] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to Complement Humans. 2020.

[305] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, pages 1920–1953, 2017.

[306] Carl O Word, Mark P Zanna, and Joel Cooper. The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of experimental social psychology*, 10(2):109–120, 1974.

[307] Eric P Xing, Michael I Jordan, Stuart J Russell, and Andrew Y Ng. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 521–528, 2003.

[308] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. FairGAN: Fairness-aware Generative Adversarial Networks. *arXiv preprint arXiv:1805.11202*, 2018.

[309] Yan Yan, Romer Rosales, Glenn Fung, and Jennifer G Dy. Active learning from crowds. In *International Conference of Machine Learning*, 2011.

[310] Chunlei Yang, Jialie Shen, and Jianping Fan. Effective summarization of large-scale web images. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1145–1148, 2011.

[311] Hongliang Yu, Zhi-Hong Deng, Yunlun Yang, and Tao Xiong. A joint optimization model for image summarization based on image content and tags. In *Twenty-eighth AAAI conference on artificial intelligence*, 2014.

[312] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017.

[313] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the Type and Target of Offensive Posts in Social Media. 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 1415–1420, 2019.

[314] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

[315] Dirk A Zetzsche, Douglas W Arner, Ross P Buckley, and Brian Tang. Artificial Intelligence in Finance: Putting the Human in the Loop. 2020.

[316] 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.

[317] Kaipeng Zhang, Lianzhi Tan, Zhifeng Li, and Yu Qiao. Gender and smile classification using deep convolutional neural networks. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 34–38, 2016.

[318] Xueru Zhang, Mohammad Mahdi Khalili, and Mingyan Liu. Long-Term Impacts of Fair Machine Learning. *Ergonomics in Design*, page 1064804619884160, 2019.

[319] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020.

[320] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, 2017.

[321] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang Chang. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[322] Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. Searching for Effective Neural Extractive Summarization: What Works and What's Next. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1049–1058, 2019.

[323] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Extractive Summarization as Text Matching. *arXiv preprint arXiv:2004.08795*, 2020.

[324] Jing Zhou, Wei Li, Jiaxin Wang, Shuai Ding, and Chengyi Xia. Default predic-

tion in P2P lending from high-dimensional data based on machine learning. *Physica A: Statistical Mechanics and its Applications*, 534:122370, 2019.

[325] Zhi-Hua Zhou and Xu-Ying Liu. On multi-class cost-sensitive learning. *Computational Intelligence*, 26(3):232–257, 2010.

[326] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.

Appendix A

Appendices

A.1 Appendix for Chapter 3

A.1.1 Implementation Details

Details of *SS-ST* baseline. The complete implementation of the semi-supervised self-training baseline *SS-ST* is given in Algorithm 5. We use $k = 5$ for PPB-2017 simulations.

Algorithm 5 *SS-ST* baseline

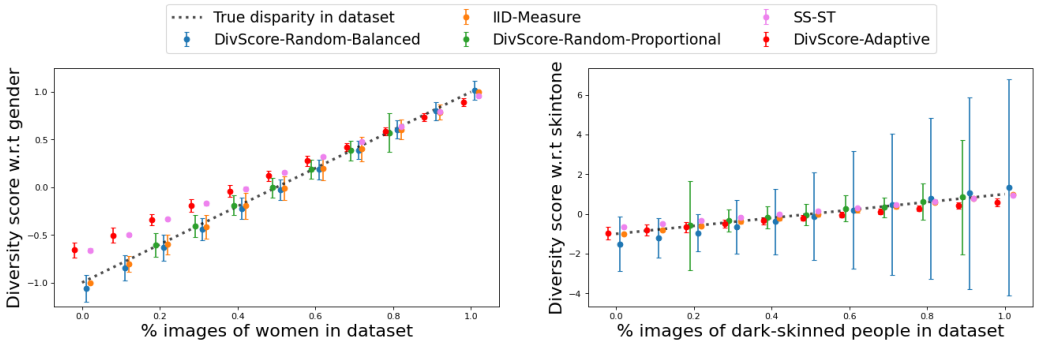
Input: Dataset S, control set $T := T_0 \cup T_1$, $\text{sim}(\cdot, \cdot)$, $k \in \mathbb{Z}_{>0}$

$n_0, n_1 \leftarrow 0$
while $S \neq \emptyset$ **do**
  for $x \in S$ **do**
    $s(x) \leftarrow \frac{1}{|T_0|} \sum_{y \in T_0} \text{sim}(x, y) - \frac{1}{|T_1|} \sum_{y \in T_1} \text{sim}(x, y)$
  $\tilde{T} \leftarrow$ top k elements in set $\{|s(x)|\}_{x \in S}$
  $n_0 \leftarrow n_0 + |\{s(x) \mid x \in \tilde{T}, s(x) > 0\}|$
  $n_1 \leftarrow n_1 + |\{s(x) \mid x \in \tilde{T}, s(x) < 0\}|$
  $S \leftarrow S \setminus \tilde{T}, T \leftarrow T \cup \tilde{T}$
return $(n_0 - n_1)/|S|$

PPB-2017 and CelebA datasets. For both PPB-2017 and CelebA datasets, feature extraction for images is done using the pre-trained VGG-16 deep network [276]. The network has been pre-trained on the Imagenet [85] dataset. To extract the feature of any given image, we pass it as input to the network and extract the 4096-dimensional weight vector of the last fully connected layer. We further reduce the feature vector size to 300 by performing PCA on the set of features of all images in the dataset.

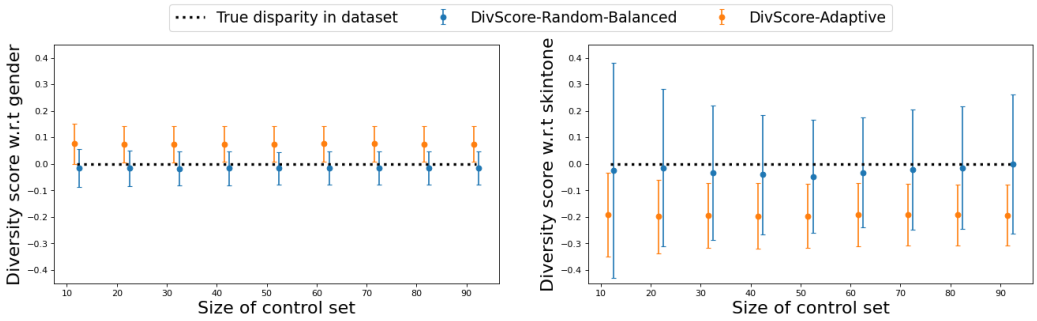
TwitterAAE dataset. For the TwitterAAE dataset, the authors constructed a demographic language identification model to report the probability of each post being written by a user of any of the following population categories: non-Hispanic Whites, non-Hispanic Blacks, Hispanics, and Asians. We filter the dataset to contain only posts for which the probability of belonging to the non-Hispanic African-American English language model or non-Hispanic White English language model is ≥ 0.99. This leads to a dataset of around 1.2 million tweets, with around 100k posts belonging to the non-Hispanic African-American English language model and 1.06 million posts belonging to the non-Hispanic White English language model; we will refer to the two groups of posts as AAE and WHE posts.

To extract feature vectors corresponding to the Twitter posts, we use a Word2Vec model [210] pre-trained on 400 million Twitter posts [121]. For any given post, we first use the Word2Vec model to extract features for every word in the post. Then we take the average of the word features to obtain the feature of the post.



(a) Gender protected attribute (b) Skin-tone protected attribute

Figure A.1: Results for PPB-2017 dataset using random and adaptive control sets. The plots in this figure are the same as the plots in Figure 3.1, except that we don't put y-axis limitations here to present the complete errorbars for all methods.



(a) Gender protected attribute (b) Skin-tone protected attribute

Figure A.2: Results for PPB-2017 dataset using different sized random and adaptive control sets.

A.1.2 Other Empirical Results

Alternate Figure 3.1 plot. First, we present the plots from Figure 3.1 without y-axis limitations. This is presented in Figure A.1.

Variation of performance with the control set size for PPB-dataset. Figure A.2 presents the variation of disparity measure with the control set size. The disparity in the collection is fixed to be 0. The plots show that *DivScore-Adaptive* can achieve low approximation error using smaller sized control sets than *DivScore-Random-*

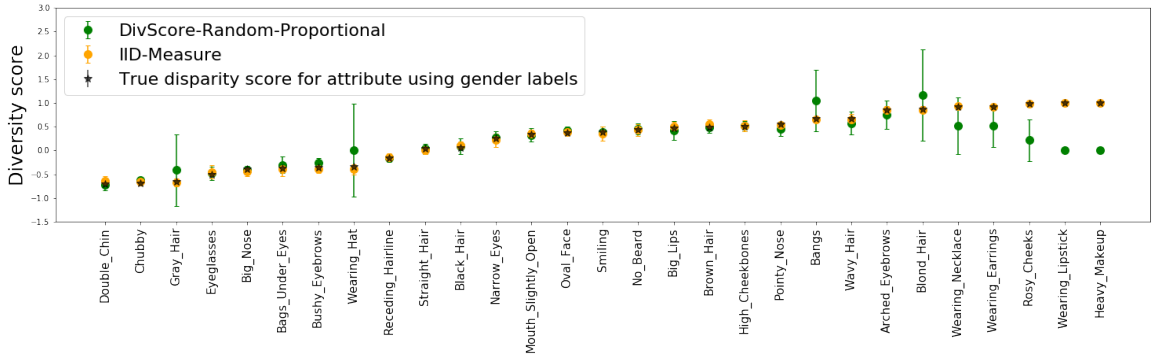


Figure A.3: Performance of *DivScore-Random-Proportional* and *IID-Measure* on CelebA dataset

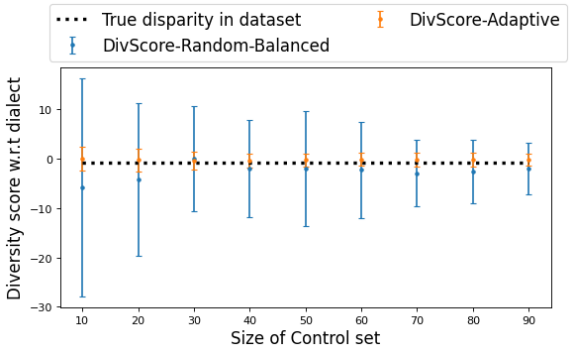


Figure A.4: Results for TwitterAAE dataset using different sized random and adaptive control sets.

Balanced.

Performance of *DivScore-Random-Proportional* and *IID-Measure* on CelebA dataset. Figure A.3 presents the performance of *DivScore-Random-Proportional* and *IID-Measure* for different facial attributes of the CelebA dataset. As expected, *IID-Measure* has a low approximation error, while *DivScore-Random-Proportional* has a low approximation error for some attributes and a high error for others. Nevertheless, as discussed in Section 3.4.2, both baselines need different control sets for collections corresponding to different attributes, and hence, are costly when auditing multiple collections from the same domain.

Variation of performance with the control set size for TwitterAAE-dataset. Figure A.4 presents the variation of disparity measure with the control set size. The disparity in the collection is fixed to be -0.826 (which is the disparity of the overall dataset) The plots show that, once again, *DivScore-Adaptive* can achieve low approximation error using much smaller sized control sets than *DivScore-Random-Balanced*.

A.2 Appendix for Chapter 4

A.2.1 Details of baselines

In this section, we provide the details of the baselines against which compare our algorithms. The first is determinant-based diversification [177, 52], *DET*. This approach effectively diversifies the selected images across their feature space. Suppose that we need to return M images corresponding to the query q. Given the query similarity scores $A(q, x)$, we can sort the list in ascending order and extract the first $c \cdot M$ images from the list, where $c > 1$ (we use $c \approx 3$ in our experiments), denoted by $\mathcal{W}_{c,q}$. We can then employ the following standard diversification technique to find the most diverse images in the set $\mathcal{W}_{c,q}$. For any $W \subseteq \mathcal{W}_{c,q}$, such that $|W| = M$, let V_W denote the matrix with the feature vectors of images in W as rows. Then return the set

$$\arg \max_{W \subseteq \mathcal{W}_{c,q}} \det\left(V_W V_W^\top\right).$$

If the number of subsets W is large (can be exponential), we use greedy approximate algorithms for this task [227].

Next, we compare with respect to another algorithm that aims to reduce redundancy in the final set, *MMR*. The algorithm is an iterative algorithm that starts with an empty set R and adds one image to R in each iteration. The chosen image is the one that minimizes the score

$$\alpha \cdot A(q, x) - (1 - \alpha) \cdot \min_{x' \in R} \sim (x, x').$$

The first part of the above expression captures query relevance while the second part penalizes an image according to similarity to existing images in the summary R. This algorithm (also referred to as *maximum marginal relevance*) is a popular

document summarization algorithm to reduce redundancy [46]. We will use $\alpha = 0.5$.

The baselines *DET* and *MMR* aim to show the importance of having a control set. In the absence of any attribute information with respect to which the results are expected to be diverse (for example, say gender), directly diversifying the output images will result in images that are diverse in unimportant features like background. The control set T helps us identify the features for which diversity should be ensured.

For the third and fourth baseline, we will use automatic gender classification tools. Using existing pre-trained gender classification models, in particular, [188] [1], we derive the gender labels for the images in the small dataset.

The third baseline, *AUTOLABEL*, is the following: we select $M/2$ images labeled *male* (by the classification tool) with the best query relevance score $A(q, x)$ and $M/2$ images labeled *female* with the best query relevance score $A(q, x)$. For evaluation, however, we use the true gender labels of the images. The purpose of this baseline is to show that using existing *imperfect* auto-labeling tools to set constraints for diversification can lead to magnifying the biases already present in the pre-trained classification model used.

For the fourth baseline *AUTOLABEL-RWD*, we use the monotone submodular function proposed by [193]. They suggest that instead of penalizing a subset for having redundant images, one should reward a subset for being diverse. The scoring function to measure the quality of a set R is then the following (adapted for our domain):

$$\text{rwd}(R) := \sum_{x \in R} A(q, x) + \sum_{i=1}^{K} \sqrt{\sum_{x \in R \cap P_i} A(q, x)},$$

where $P_1, \ldots, P_K$ are the partitions of the domain based on the protected attribute. For the case of gender, we will have two partitions. The second part of the expres-

[1] https://github.com/dpressel/rude-carnie

sion ensures that adding images from different partitions has a higher diversity score than adding images from the same partition. Once again, we will create the partitions according to the gender labels obtained using the classification tool. We will use a greedy algorithm to obtain an approximately optimal subset for this case, since finding the optimal solution directly has a large time complexity. The greedy algorithm will simply add the image $\arg\min_{x \in S \setminus R} \text{rwd}(R \cup \{x\})$ at every step, where R is the subset chosen so far.

A.2.2 Implementation details

In this section, we provide the complete implementation details, starting with the query matching algorithm $A(\cdot, \cdot)$ and the similarity function $\text{sim}(\cdot, \cdot)$ used in our empirical analysis.

Image similarity

To obtain the similarity score $\text{sim}(x_1, x_2)$ for two given images, we can utilize a pre-trained convolutional neural network. We use the VGG-16 network [276], a 16-layer CNN, pre-trained on Imagenet [85] dataset, for generating the feature vectors[2]. We take the weights of the edges from the last fully-connected layer as the feature vector for the image. The process can be summarized in the following steps [3] [4]: (1) feed the image x_1, x_2 into the VGG-16 network and obtain the feature vectors v_{x_1}, v_{x_2} of dimension 4096, (2) perform Principal Component Analysis to reduce the feature vector size, (3) return the cosine distance as similarity score, i.e.,

$$\text{sim}(I_1, I_2) = 1 - \frac{v_{I_1} \cdot v_{I_2}}{\|v_{I_1}\|_2 \|v_{I_2}\|_2}.$$

[2] other networks such as [270] could similarly be used instead.

[3] Similar to one-shot learning using Siamese Networks [174].

[4] Cosine distance has been used in document summarization literature to calculate similarity [193] as well. The cosine distance metric also outperformed other norm-based metrics, such as 1-norm.

This method of using pre-trained models for other tasks is also called "transfer learning". This technique has been successfully employed in many other image-related tasks [238].

Query matching

QS-balanced (Algorithm 4) and *MMR-balanced* use a black-box querying algorithm A to rank images according to similarity to a query. For evaluation purposes, we describe an algorithm for query matching algorithm in the case of Occupations and CelebA datasets.

Query matching algorithm A for Occupations dataset. Suppose that for every q, we are provided a small set of images T_q; for example, for query "doctor", 10 images of doctors (that can be hand-verified). Then using $\sim$ function, for the query set T_q and for each image $x \in S$, we can calculate the score $\text{avgSim}_{T_q}(x) := \text{avg}_{x' \in T_q} \text{sim}(x, x')$. The score $\text{avgSim}_{T_q}(x)$ gives us a quantification of how similar the image x is to all other images in set T_q, and correspondingly how similar it is to query q. Before using this score further, we can normalize it by subtracting the mean and dividing by standard deviation. Therefore given a set T_q, for each $x \in S$, the query similarity score can be defined as

$$A(q, x) := \widehat{\text{avgSim}}_{T_q}(x) = \frac{\text{avgSim}_{T_q}(x) - \text{mean}(\text{avgSim}_{T_q})}{\text{std}(\text{avgSim}_{T_q})}.$$

We will use this score to compute $\text{DS}_q()$.

For each query occupation q, we use the top 10 images from Google results of that occupation in the dataset as the similarity control set T_q. Note that we use this query relevance algorithm for other baselines which employ $A(q, \cdot)$ score as well.

When we have to report accuracy for results over the Occupations dataset, we

will use the measure of query similarity. While the above score $\widehat{\text{avgSim}}_{T_q}$ is a measure of query similarity, it represents high similarity if the value is lower. To avoid confusion, and maintain the convention that a high value is a high accuracy when measuring accuracy we will use $\text{sim}(x_1, x_2) = \frac{v_{x_1} \cdot v_{x_2}}{\|v_{x_1}\|_2 \|v_{x_2}\|_2}$ for this chapter and then similarity calculate average similarity with respect to all query images.

Query matching algorithm A for CelebA dataset For the CelebA dataset, recall that we divide the dataset into train and test partitions. The train partition is used to train a multi-class classification model, with the facial attributes as the labels.

The classification model, given an input image, returns a vector of length 37, where each entry ($\in [0,1]$) represents the probability that the input image satisfies the corresponding attribute; let $f : S \to [0,1]^{37}$ denote the classifier. We use the MobileNetV2 architecture and a transfer learning approach suggested by Anzalone et al. [14] for the classifier, which achieves a training accuracy of around 90%.

Since we follow the convention that the smaller the score the better the image corresponds to the query, we will use the negative of the classifier output as the query-similarity score, i.e., $A(q, x) = -f^{(q)}(x)$, where $f(x)$ denotes the output of classifier for image x and $f^{(q)}$ denotes the entry corresponding to the attribute q.

For the image-similarity scoring function, we will use the pre-trained VGG-16 network to extract the features of the images and return the cosine distance between the features as the similarity score between the images.

Diversity Control Matrix

Finally, to efficiently implement *QS-balanced*, we can construct a diversity control matrix of size $|S| \times |T|$ using the image-similarity scores between the images in S and images in T. Before using this matrix to compute the DS_q scores, we will

normalize each column of this matrix, i.e., we compute $\widehat{\text{avgSim}}_{\{x_c\}}(x)$. Therefore the final $\text{DS}_q(x, x_c)$ score is evaluated as

$$\text{DS}_q(x, x_c) = \alpha \cdot \widehat{\text{avgSim}}_{\{x_c\}}(x) + (1 - \alpha) \cdot \widehat{\text{avgSim}}_{T_q}(x).$$

To implement this approach efficiently, we calculate the scores $\text{sim}(x, x_c)$ as a pre-processing step and store them in the diversity control matrix. Then given a query, we calculate the scores $\widehat{\text{avgSim}}_{T_q}(x)$ and combine the diversity control matrix and query similarity score list to get a matrix of size $|S| \times |T|$, where the element corresponding to $x \in S$ and $x_c \in T$ has the value $\text{DS}_q(x, x_c)$.

For *MMR-balanced*, we use the greedy approach and add the diversity score of an image to its relevance score at every step.

A.2.3 Model Properties

As mentioned earlier in the Related Work section, query-based diverse summarization has been a major area of research in many sub-domains within information retrieval. For diverse document and image summarization, multiple models have been considered and evaluated rigorously [46, 193, 294]. Some of the models we consider as baselines are derived from models that are popular and commonly used in diverse document summarization literature (*MMR*, *DET* and *AUTOLABEL-RWD*). One of the properties that a lot of diversity-ensuring summarization models share is the property of submodularity, defined formally below.

Definition A.2.1 (Submodular function)**.** *Given a set of elements* $\Omega = \{x_1, \ldots, x_n\}$ *and a function* $f : 2^{\Omega} \to \mathbb{R}$*, the function* f *is called submodular if it satisfies the property that for any* $R_1 \subseteq R_2 \subseteq \Omega$ *and any element* $x \in \Omega$*,*

$$f(R_1 \cup \{x\}) - f(R_1) \geq f(R_2 \cup \{x\}) - f(R_2).$$

Submodular functions quantify the property of *diminishing returns*, and in many settings, a simple greedy approach can return a good approximation of the optimal solution for maximizing a submodular function. In the case of maximizing monotone submodular functions subject to cardinality/matroid constraints, a greedy algorithm returns a 0.632-factor approximation to the optimal subset in the worst case [227, 43], and in many cases, performs much better than the worst-case bound.

Submodular functions occur naturally when the task is to ensure that the output summary is representative of a particular subdomain of the population. For example, in the case of image summarization tasks that aim to reduce redundancy or ensure representativeness in the final set, Tschiatschek et al. [294] argued that many models in existing literature are cases of submodular maximization. Even algorithms based on determinantal point processes, such as [52, 177] satisfy this property since the determinant-based objective function is log-submodular.

In this section, we show that the scoring mechanisms considered in Chapter 4 satisfy the diminishing returns property and are in line with the submodularity property common to the diverse summarization literature. The submodularity of *MMR*, *AUTOLABEL-RWD* [193] and *DET* [177] has been already discussed and proved in multiple prior works. We primarily focus on the *QS-balanced* and *MMR-balanced* algorithms.

Reducing Redundancy. A simple algorithm to reduce redundancy in the output summary is the following: let R denote summary; at each step add the image $x \in S \setminus R$ which minimizes the following score

$$\alpha \cdot A(q, x) - (1 - \alpha) \cdot \min_{x' \in R} \text{sim}(x, x'), \tag{3}$$

where $\alpha \in [0, 1]$. We use this expression, called the maximum marginal relevance, as a baseline in our experiments as well, and it is common in document summarization algorithm [46, 193].

While this expression ensures the images are visibly diverse, it cannot focus on the features with respect to which diversity is desired by the user (as seen in Section 4.4.3). For example, it may ensure that the images in the summary have very different backgrounds but cannot ensure the gender proportion of the people in the image summary is equal. This leads us to use a control set.

Diversity using a control set. To ensure visible diversity in the results we use a control set T. Adding the control set similarity score to expression (3), we get the following relevance score for adding an image x to a set R,

$$\text{mmod}_R(x) := (1 - \alpha - \beta) \cdot A(q, x) + \alpha \cdot \min_{x_c \in T} \text{sim}(x, x_c) - \beta \cdot \min_{x' \in R} \text{sim}(x, x'), \quad (4)$$

where $\alpha, \beta \in [0, 1]$. The second term in the above expression now also aims to find the image in the control set T most similar to x. If an image corresponding to x_c has already been chosen, call it x', and x has a large similarity with x_c as well, then we don't want to choose x. In this case values $\text{sim}(x, x_c)$ and $\text{sim}(x, x')$ will be close and partially cancel each other, ensuring that the overall expression doesn't have the minimum value.

Recall that we use this scoring function as a baseline in our experiments as well. Furthermore, the expression (4) satisfies the diminishing-returns property.

Lemma A.2.1 (Submodularity of (4))**.** *Let $f : 2^S \to \mathbb{R}$ be a function such that $f(R \cup \{x\}) - f(R) = -mmod_R(x)$. Then f is submodular.*

Proof. For each x, $(\alpha \cdot A(q, x) + \beta \cdot \min_{x_c \in T} \text{sim}(x, x_c))$ is constant and independent

of the set R. Consider two subsets $R_1 \subseteq R_2$. Then

$$\min_{x' \in R_1} \text{sim}(x, x') \geq \min_{x' \in R_2} \text{sim}(x, x'),$$

since the chance of an image in R_2 being similar to x is larger than that for R_1. Correspondingly, this score satisfies the diminishing-returns property. □

Alternate summarization relevance expression. Note that while the above algorithm can ensure diversity and non-redundancy, it has two major problems.

The first problem is that in the presence of a control set, the primary aim is to ensure diversity in the output set of images with respect to the features in the control set, and not the overall feature space. For such a task, the score $\min_{x' \in R} \text{sim}(x, x')$ may not ensure complete diversity with respect to the features of the control set due to the additional goal of reducing redundancy. This was also observed in the empirical results presented in Section 4.4.3; the standard deviation of the fraction of women in top results was higher for *MMR-balanced* results compared to *QS-balanced* results. Hence we can try to slightly relax the goal of reducing redundancy to ensure better diversity with respect to control set features.

The second problem is the time complexity. The iterative algorithm, based on choosing the image with the lowest score according to (2), is very slow. This is due to the fact that it has to evaluate the non-redundancy score $\min_{x' \in R} \text{sim}(x, x')$ at each step of the algorithm. Once again, we can instead use T directly to ensure diversity and reduce the time complexity.

This leads us to our main algorithm, which addresses both of these issues. Given the parameter $\alpha \in [0, 1]$ and a query q, for each $x_c \in T$, our primary scoring

function $\mathrm{DS}_q(\cdot,\cdot)$ is the following:

$$\mathrm{DS}_q(x, x_c) = \alpha \cdot \mathrm{sim}(x, x_c) + (1-\alpha) \cdot A(q, x).$$

We can show that this algorithm also corresponds to the diminishing returns property. Furthermore, since it does not include any term to reduce redundancy by checking already chosen elements, using appropriate pre-processing (as mentioned in Section 4.4) it is much faster than *MMR-balanced*.

Diminishing-returns property of *QS-balanced* (Algorithm 4). To show that *QS-balanced* also satisfies the diminishing returns property, we will present an alternative iterative algorithm that outputs the same set as *QS-balanced* (Algorithm 4). For simplicity, assume that the size of the desired summary is a multiple of $|T|$. Let $U : T \to 2^S$, be the following function $U(x_c) := \left\{x \in S \mid x_c = \arg\min_{x'_c \in T} \mathrm{sim}(x, x'_c)\right\}$. Consider an iterative algorithm that adds one image to the final subset R in each iteration. The image is chosen according to the following score function:

$$\mathrm{DDS}_R(x) := \begin{cases} \frac{u}{2^n}, & \text{if } \exists x_{c_1}, x_{c_2} \in T, \\ & \text{s.t., } x_{c_1} \neq x_{c_2}, x = \arg\min_{x' \in U(x_{c_1}) \setminus R} \mathrm{DS}_q(x, x_{c_1}), \\ & |U(x_{c_1}) \cap R| = n, \text{ and} \\ & |U(x_{c_2}) \cap R| > n \\ \frac{l}{2^n}, & \text{otherwise,} \end{cases} \tag{5}$$

where $U(x_c)$ is as defined earlier and $u, l \in \mathbb{R}$ are numbers such that $l < u \leq 2l$. Then we can prove the following theorems about this expression.

Theorem A.2.2. *Given a dataset S, control set T, query q, query relevance algorithm A and numbers u, l, such that $l < u \leq 2l$, the set returned by Algorithm 4 is the same as the*

set returned by the iterative algorithm using the scoring function (5).

Proof. As mentioned earlier (Figure 6.1), Algorithm 4 is based on constructing a $|T| \times |S|$ matrix using scores $\mathrm{DS}_q(x, x_c)$, and then sorting each row of the matrix. The images are finally chosen by taking images first from the first column, then the second column, and so on. DDS score creates a similar ordering.

The first image was chosen for any $x_c \in T$ since for all of them $|U(x_c) \cap R| = 0$. The image chosen will be the image that will have the best score with respect to x_c, i.e., $x = \arg\min_{x' \in U(x_{c_1})} \mathrm{DS}_q(x', x_{c_1})$. This corresponds to Step 9 of Algorithm 4. Now $|U(x_c) \cap R| = 1$ and for all other $x'_c \neq x_c$, $|U(x_c) \cap R| = 0$, the iterative algorithm will next choose an image corresponding to a different $x'_c \neq x_c$ (since $u > l$), thus enforcing the loop in Step 8 of Algorithm 4.

Once one image is chosen for each x_c, the counter n will increase and the same process will be repeated. Since we assumed that the size of the chosen subset is a multiple of $|T|$, the ordering in which each x_c is addressed does not matter. □

Note that the above expression can be modified for the case when the size of the desired summary is not a multiple of $|T|$. To do so, one just has to fix an ordering for x_{c_1}, x_{c_2} according to the scores $\left\{\min_{x \in U(x_c) \setminus R} \mathrm{DS}_q(x, x_c)\right\}_{x_c}$.

Having established the above equivalence, we can also show that the expression satisfies the diminishing-returns property.

Lemma A.2.3 (Submodularity of (5))**.** *Let $f : 2^S \to \mathbb{R}$ be a function such that $f(R \cup \{x\}) - f(R) = DDS_R(x)$. Then f is submodular.*

The fact that the above function is submodular is in line with other functions considered for diverse image summarization, for example [52, 294].

Proof. Consider two subsets $R_1 \subseteq R_2$. Let $n_1 := \lfloor |R_1|/|T| \rfloor$ and $n_2 := \lfloor |R_2|/|T| \rfloor$. Assume that $x \in U(x_c)$. There are two cases that we need to address.

Case 1, $n_1 = n_2$. In this case, if $x = \arg\min_{x' \in U(x_c) \setminus R_2} \mathrm{DS}_q(x', x_c)$, then $\mathrm{DDS}_{R_1}(x) = \mathrm{DDS}_{R_2}(x) = u/2^{n_1}$. If x does not satisfy this condition and is not the image with the best score for x_c in this iteration, then $\mathrm{DDS}_{R_1}(x) = \mathrm{DDS}_{R_2}(x) = l/2^{n_1}$. For both cases, the score of x is equal for R_1 and R_2.

Case 2, $n_1 < n_2$. In this case, there are two sub-cases.
Either $x \neq \arg\min_{x' \in U(x_c) \setminus R_1} \mathrm{DS}_q(x', x_c)$ and $x \neq \arg\min_{x' \in U(x_c) \setminus R_2} \mathrm{DS}_q(x', x_c)$. Then $\mathrm{DDS}_{R_1}(x) = l/2^{n_1}$ and $\mathrm{DDS}_{R_2}(x) = l/2^{n_2}$. Since $n_1 < n_2$, we have that

$$\frac{l}{2^{n_1}} > \frac{l}{2^{n_2}} \implies \mathrm{DDS}_{R_1}(x) \geq \mathrm{DDS}_{R_2}(x).$$

The other sub-case is that $x \neq \arg\min_{x' \in U(x_c) \setminus R_1} \mathrm{DS}_q(x', x_c)$ and
$x = \arg\min_{x' \in U(x_c) \setminus R_2} \mathrm{DS}_q(x', x_c)$. Then $\mathrm{DDS}_{R_1}(x) = l/2^{n_1}$
and $\mathrm{DDS}_{R_2}(x) = u/2^{n_2}$. Since $n_1 \leq n_2 - 1$ and $u < 2l$, we have that

$$\mathrm{DDS}_{R_1}(x) = \frac{l}{2^{n_1}} > \frac{u}{2^{n_1+1}} \geq \frac{u}{2^{n_2}} = \mathrm{DDS}_{R_2}(x).$$

Hence the score DDS follows the diminishing returns property for all cases. □

A.2.4 Additional Empirical Results on Occupations Dataset

In this section, we present additional details and empirical results for the Occupations dataset. While the results in the main body represent the best choice of parameters for the algorithm, such as α value or the control set, we also present here the empirical results corresponding to varying parameters so as to motivate the choices made for the simulations in the main body.

Control sets

For the Occupations dataset, we evaluate our approach on four different small (10-30 images) control sets in order to evaluate the effect of the control set on the end result. Two sets (Control Set-1 and Control Set-2) are hand selected by the authors using images from Google results and are intended to be diverse with respect to presented gender and skin color. The other two sets (PPB Control Set-1 and PPB Control Set-2) are generated by randomly sub-sampling from the Pilot Parliaments Benchmark Dataset [39]. This dataset has gender and skin-tone labeled images, and we select images uniformly at random conditioned on selecting an equal number of men and women and an equal number of people from all skin-tones. The control sets are presented in Figure A.5.

Intersectionality results for all algorithms

We present the detailed intersectionality comparison with all baselines in the following table. This is an extension of Table 4.1. The performance of *QS-balanced* algorithm can be observed to be better than other baselines in terms of intersectional diversity.

(a) Control Set - 1

(b) Control Set - 2

(c) PPB Control Set - 1

(d) PPB Control Set - 2

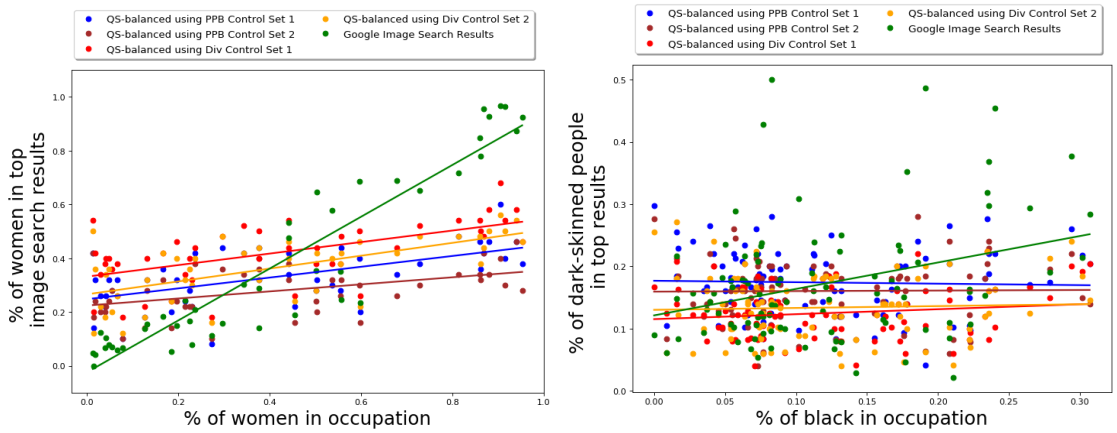
Figure A.5: Occupations dataset: Control Sets used in the experiments. The first two diversity controls (a) and (b) are hand-picked while the last two (c) and (d) were randomly sampled from the PPB dataset.

Table A.1: Occupations dataset: Intersectionality comparison with all baselines.

Algorithm	% gender stereotypical with fair skin	% gender anti-stereotypical with fair skin	% gender stereotypical with dark skin	% gender anti-stereotypical with dark skin
QS-balanced	.46 (.14)	.37 (.14)	.09 (.05)	.08 (.05)
MMR-balanced	.46 (.17)	.39 (.18)	.09 (.06)	.06 (.04)
Google	.60 (.20)	.24 (.21)	.11 (.08)	.05 (.07)
MMR	.57 (.21)	.30 (.21)	.07 (.06)	.05 (.05)
DET	.52 (.12)	.33 (.12)	.09 (.05)	.06 (.05)
AUTOLABEL	.54 (.16)	.31 (.16)	.09 (.06)	.06 (.04)
AUTOLABEL-RWD	.56 (.19)	.30 (.19)	.08 (.06)	.05 (.05)

Results for different control sets

As noted earlier, we use 4 different control sets in our empirical evaluations. The results presented in the main body correspond to the evaluation using PPB-control set 1. We provide the diversity comparison for different control sets in Figure A.6.



(a) Gender diversity comparison

(b) Skin-tone diversity comparison

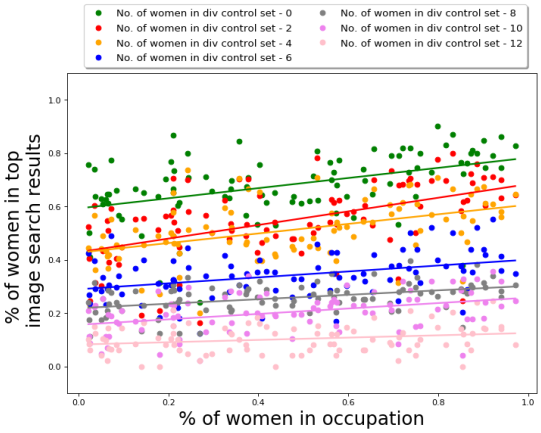
Figure A.6: Occupations dataset: Gender and skin-tone diversity comparison of results of *QS-balanced* algorithm on different control sets. For gender, using any of the control sets results in a more gender-balanced output. For skintone, using PPB Control Set-1 results in the best results among all control sets. For most occupations, the top Google images have a much larger or much smaller fraction of images of dark-skinned people.

Results for different compositions of control sets

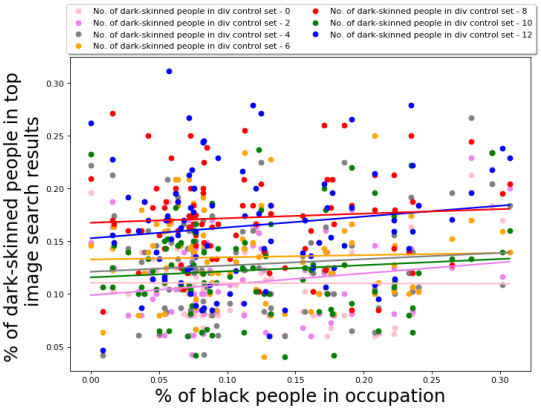
To explicitly see the impact of diversity control on the diversity of the output of the algorithm, we can vary the content of the control set and observe the corresponding changes in the results. We first vary the fraction of women in the control set. The control sets are randomly chosen for the PPB-dataset, while maintaining the desired gender ratio. The results for different control sets are presented in Figure A.7a. The figure shows that increasing the fraction of women in the control set leads to an increase in the fraction of women in the output set.

Similarly, increasing the fraction of images of dark-skinned people in the con-

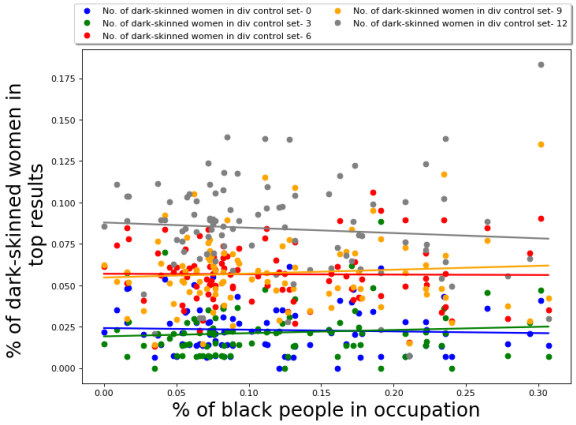
trol set leads to an increase in the fraction of images of dark-skinned people in the output; this is shown in Figure A.7b. Finally, Figure A.7c shows the impact of variation of images of dark-skinned women in the control set on the output. While the fraction of dark-skinned women still increases, it seems to be upper bounded by the fraction of images of dark-skinned women in the dataset.



(a) Different fraction of women in Control Set



(b) Different fraction of dark-skinned in Control Set



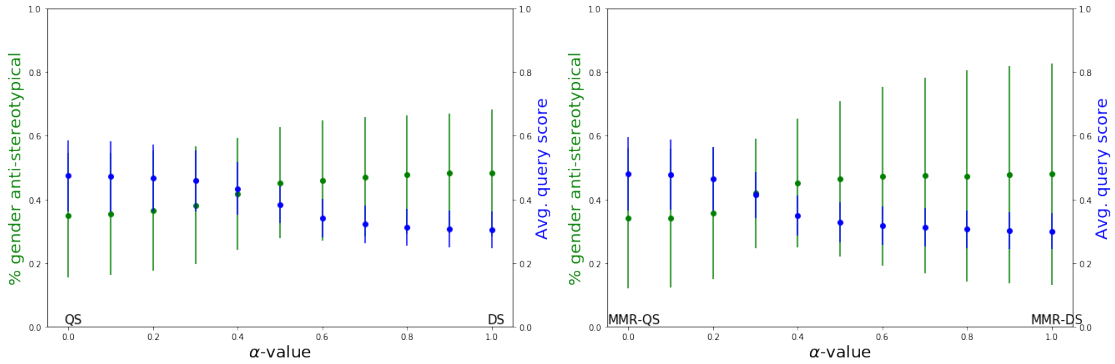
(c) Different fraction of dark-skinned women in Control Set

Figure A.7: Occupations dataset: Performance *QS-balanced* algorithm on control sets with different compositions.

Results for different α values

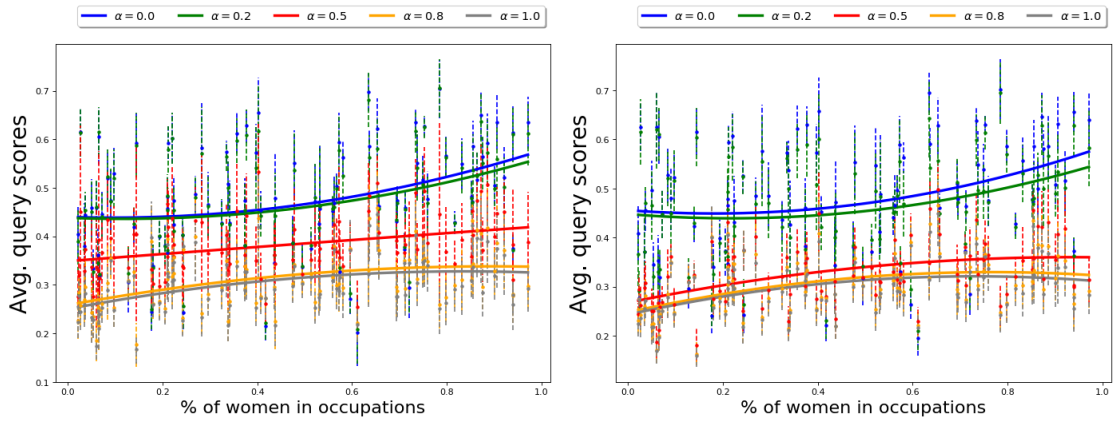
We vary the quality-fairness parameter α and look at its impact on the performance of our algorithms. The diversity results are presented in Figure A.8, while Fig-

ure A.9 expands on the accuracy for different alphas.



(a) QS-balanced (b) MMR-balanced

Figure A.8: Occupations dataset: Gender diversity and query similarity comparison of results of *QS-balanced* and *MMR-balanced* algorithms for different α-values.



(a) QS-balanced (b) MMR-balanced

Figure A.9: Occupations dataset: Accuracy of results of *QS-balanced* and *MMR-balanced* algorithms for different α-values.

For both *QS-balanced* and *MMR-balanced*, the fraction of gender anti-stereotypical images increases as the α value increases. With an increase in fairness, a loss in accuracy is expected. While the figure shows a small change in average query scores, the standard deviation of the scores seem to be decreasing as well, showing that as α increases, the dependence on the query decreases. Hence a balance between

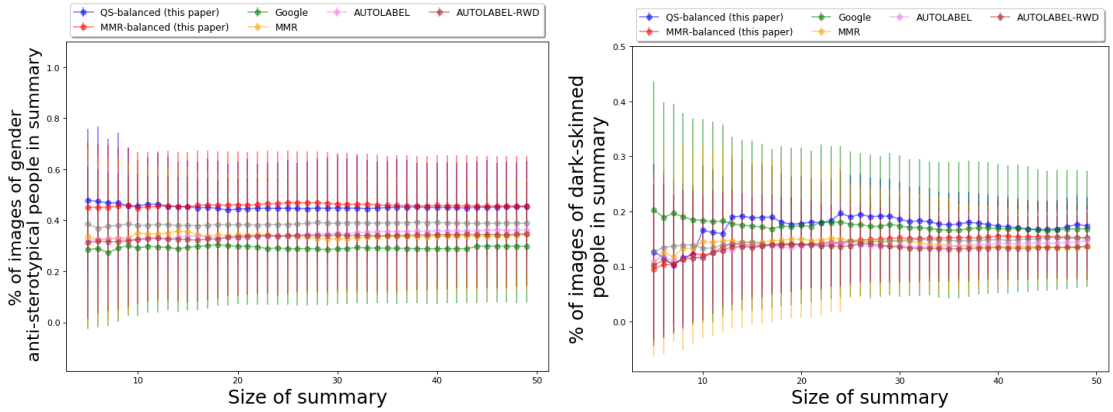
query similarity and diversity score has to be maintained by choosing an appropriate value of α, such as 0.5.

Results for different summary sizes

While the results we have presented so have been with respect to a summary of size 50. However, the size of the summary can depend on the application and the results in the first page of any web-search application will depend on the size of the screen or the device being used. Correspondingly, it is important to analyze the results for different summary sizes as well.

For *QS-balanced*, *MMR-balanced* and the baselines, we look at the average fraction of images of gender anti-stereotypical and dark-skinned people in the top k results, where k ranges from 2 to 50; the average is taken over all occupations. The results are presented in Figure A.10. We also present the gender and skintone-diversity comparison of our method vs baselines for summary sizes 10 and 20 in Figure A.11 and Figure A.12.

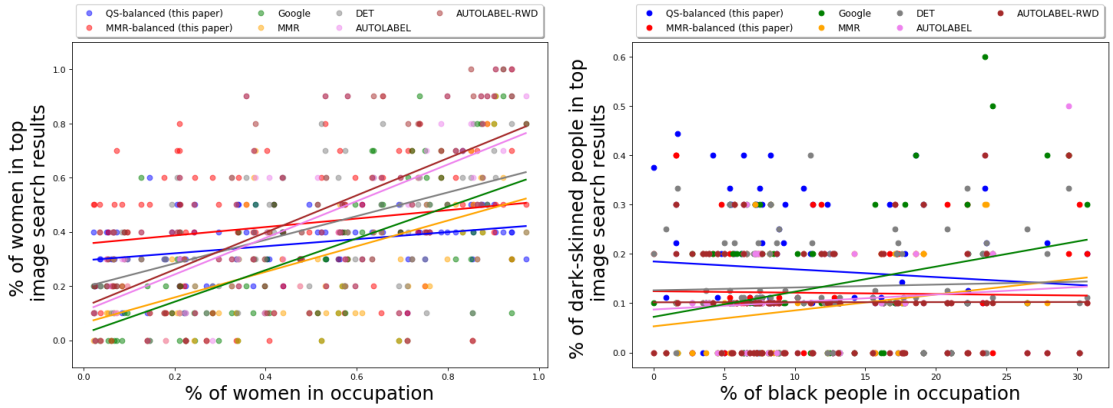
The figures shows that *QS-balanced* and *MMR-balanced* return a larger fraction of gender anti-stereotypical images for all summary sizes. With respect to skintype, Google results seem to have a larger value for average fraction of dark-skinned people for smaller summary sizes; however, the performance of *QS-balanced* in this respect is similar to better for larger summary sizes. Furthermore, Google results also have a significantly larger standard deviation, implying that the fraction of dark-skinned people is also much lower than average for some occupations.



(a) Fraction of images of gender anti-stereotypical people vs summary size

(b) Fraction of images of dark-skinned people vs summary size

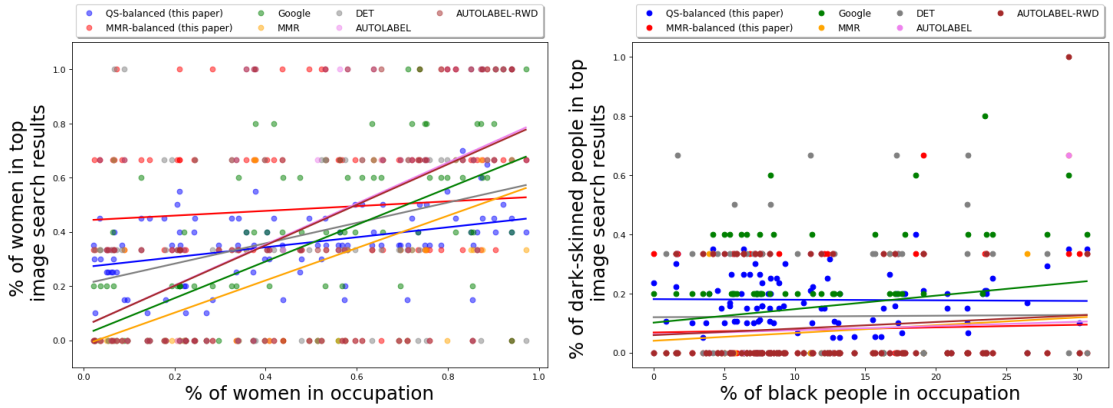
Figure A.10: Occupations dataset: Variation of the fraction of anti-stereotypical images vs size of summary for all algorithms.



(a) Fraction of images of gender anti-stereotypical people vs ground truth

(b) Fraction of images of dark-skinned people vs ground truth

Figure A.11: Occupations dataset: Fraction of anti-stereotypical images for summary size 10.



(a) Fraction of images of gender anti-stereotypical people vs ground truth

(b) Fraction of images of dark-skinned people vs ground truth

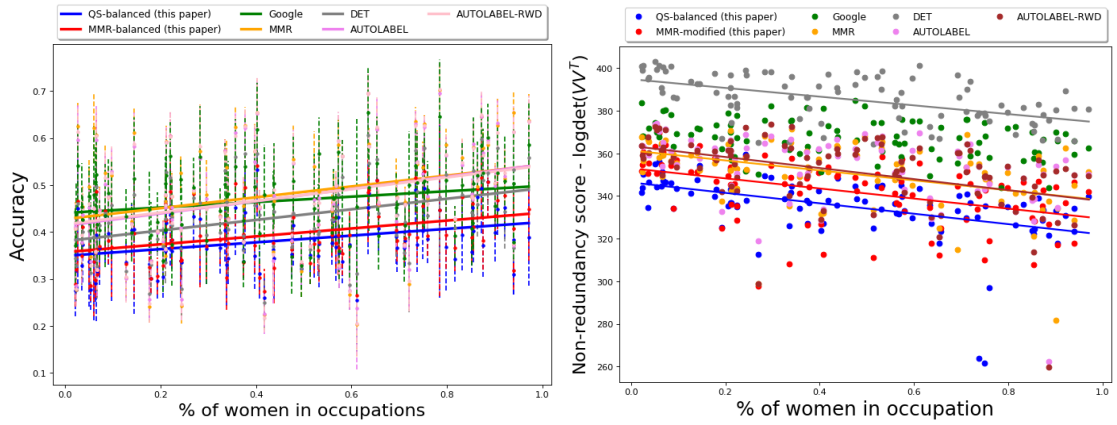
Figure A.12: Occupations dataset: Fraction of anti-stereotypical images for summary size 20.

Similarity and Non-redundancy comparison for Occupations dataset

As mentioned earlier, the accuracy for the Occupations dataset is measured using average query similarity, i.e., similarity to the set of images corresponding to the given query. We present the accuracy comparison for our methods and baselines in Figure A.13. The accuracy score of the results of all algorithms is close to each other, showing that using control set does not adversely impact the accuracy.

The second figure also presents the non-redundancy comparison of our methods and baselines. The non-redundancy measure used is the log of the determinant of the feature kernel matrix, i.e., if for a summary S, if V_S is the matrix with columns representing the feature vectors of the images in S, then the non-redundancy is measured as $\log\det\left(V_S V_S^\top\right)$ (the determinant can be pretty large and computationally more difficult to calculate, hence the logarithm). As expected, the results from *DET* have the largest non-redundancy score. The non-redundancy scores of *QS-balanced* and *MMR-balanced* are the lowest, perhaps due to enforcing fairness constraints using the control set. However, as we saw earlier, non-redundancy

does not imply diversity with respect to protected attributes.



(a) Avg. query scores comparison

(b) Non-redundancy scores comparison

Figure A.13: Occupations dataset: (a) Comparison of accuracy, as measured using mean query similarity scores, of top 50 results across all occupations. For each occupation, we also plot the mean similarity to the query control set and the standard deviation using the dotted lines. The mean similarity score of the results of all algorithms is close to each other, showing that using a control set does not adversely impact the accuracy. (b) Comparison of non-redundancy scores. As expected, the results from *DET* have the largest non-redundancy score, measured as the log of the determinant of the product of the feature matrix the output images and its transpose. The non-redundancy scores of *QS-balanced* and *MMR-balanced* are the lowest, perhaps due to enforcing fairness constraints using the control set.

Occupation accuracy of QS-balanced algorithm

Finally, we also present the accuracy of the results of the QS-balanced algorithm. The accuracy is measured as the number of images in the summary belonging to the queried occupation. The results for this accuracy are presented in Figure A.14. Note that accuracy is not a good measure of quality in this case; this is because a lot of occupations have similar-looking images. For example, images of lawyers and financial analysts are very similar, and images of doctors and pharmacists are very similar. Hence when using image similarity as a method of query matching, one cannot expect the matched images to always belong to the same query. This problem is relatively less visible for the CelebA dataset since in that case, the query

similarity algorithm is more specialized to the dataset.

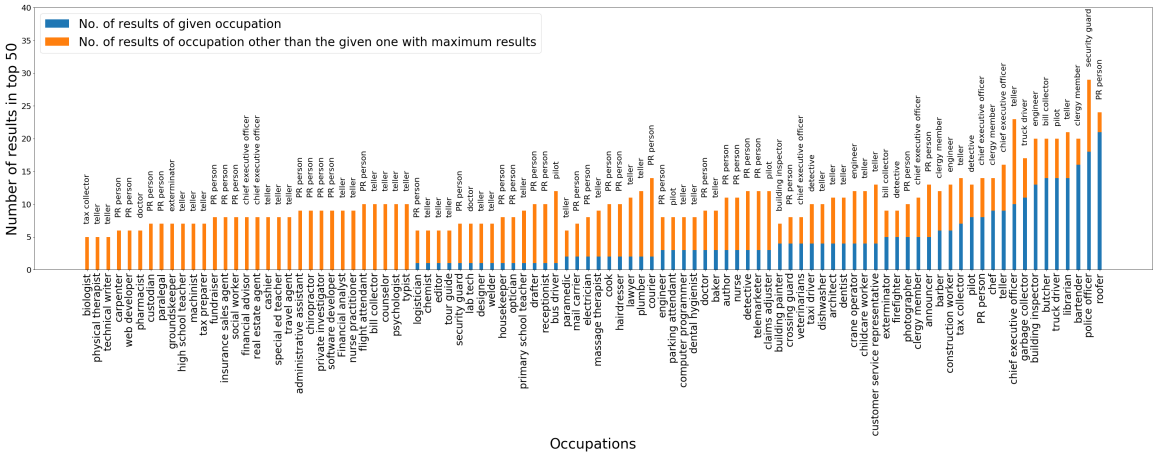


Figure A.14: Occupations dataset: Accuracy comparison of results of *QS-balanced* algorithm for different occupations. For each occupation and its summary, we present the number of images belonging to that occupation in the summary, as well as the other occupation with the highest number of images in the summary.

We also present the bar graph for when 1-norm is used, instead of cosine distance for similarity in Figure A.15. In this case, the accuracy is much worse and this is the reason for using cosine distance over 1-norm distance for all our simulations.

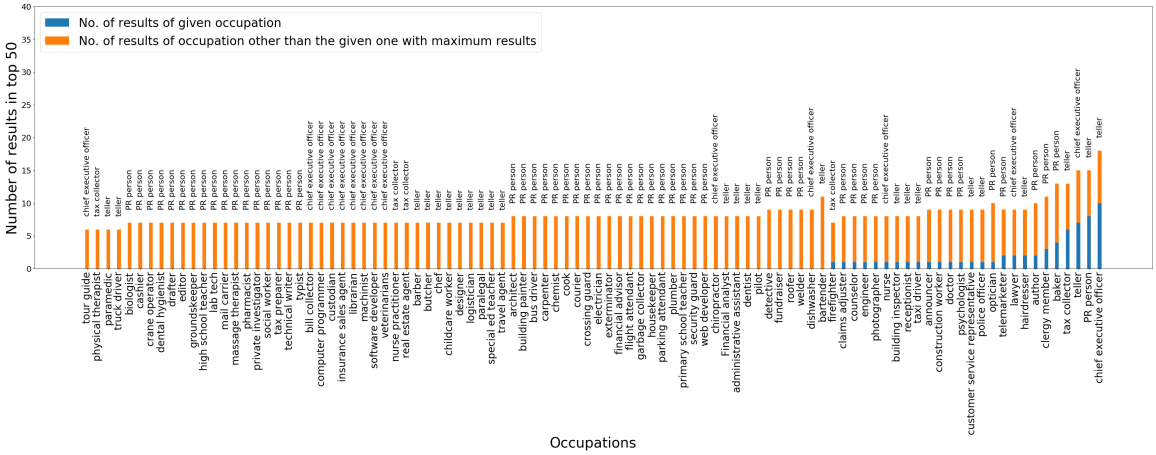


Figure A.15: Occupations dataset: Accuracy comparison of results of *QS-balanced* algorithm for different occupations using 1-norm for similarity.

Table A.2: Fraction of images with given attributes in CelebA dataset.

Attribute	Fraction of images of women with given attribute	Attribute	Fraction of images of women with given attribute
Heavy Makeup	1.0	Wearing Lipstick	0.99
Rosy Cheeks	0.98	Wearing Earrings	0.96
Blond Hair	0.94	Wearing Necklace	0.94
Arched Eyebrows	0.92	Wavy Hair	0.82
Attractive	0.77	Bangs	0.77
Pale Skin	0.76	Pointy Nose	0.76
Big Lips	0.73	High Cheekbones	0.72
No Beard	0.7	Brown Hair	0.69
Oval Face	0.68	Smiling	0.65
Mouth Slightly Open	0.63	Narrow Eyes	0.56
Blurry	0.53	Straight Hair	0.52
Black Hair	0.48	Receding Hairline	0.39
Wearing Hat	0.3	Bags Under Eyes	0.29
Bushy Eyebrows	0.28	Big Nose	0.25
Eyeglasses	0.21	Gray Hair	0.15
Chubby	0.12	Double Chin	0.12
5 o Clock Shadow	0.0	Bald	0.0
Goatee	0.0	Male	0.0
Mustache	0.0	Sideburns	0.0

A.2.5 Additional Results on CelebA Dataset

In this section, we present additional details and empirical results for the CelebA dataset. The additional results correspond to varying different parameters in the algorithm, such as α value or the control set.

Attributes of the dataset

We first present the list of facial attributes in the dataset and the fraction of images with a given attribute that are also labeled "Female" in Table A.2.

Control Sets

Once again, we will use four different control sets for our evaluation, two of them have 8 images and the other two have 24 images; the exact images are provided in Section A.2.5. The control sets are constructed by randomly sampling an equal number of images with and without the "Male" attribute from the train set. The control sets are presented in Figure A.16.

Results by features

We first present the exact gender and accuracy results by features in Figure A.17.

Results for different control sets

As noted earlier, we use 4 different control sets in our empirical evaluations. The results presented correspond to the evaluation using Control Set-4. We provide the accuracy and diversity comparison for different control sets in Figure A.22.

Results for different compositions of control sets

To explicitly see the impact of diversity control on the diversity of the output of the algorithm, we once again vary the content of the control set and observe the corresponding changes in the results. In this case, we only vary the fraction of women in the control set. The control sets are randomly chosen from the training dataset while maintaining the desired gender ratio. The results for different control sets are presented in Figure A.18. The figure shows that increasing the fraction of women in the control set leads to an increase in the fraction of women in the output set.

Non-redundancy comparison

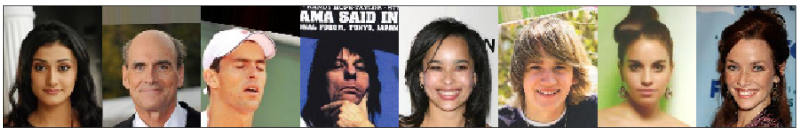
Figure A.19 presents the non-redundancy comparison of our methods and baselines. Recall that the non-redundancy measure used is the log of the determinant of the feature kernel matrix, i.e., if for a summary S, if V_S is the matrix with columns representing the feature vectors of the images in S, then the non-redundancy is measured as $\log\det\left(V_S V_S^\top\right)$. As expected, the results from *DET* have the largest non-redundancy score for most attributes. However, once again, non-redundancy does not imply diversity with respect to protected attributes.

Results for different α values

We vary the quality-fairness parameter α and look at its impact on the performance of our algorithms. The results are presented in Figure A.20. For both *QS-balanced* and *MMR-balanced*, the fraction of gender anti-stereotypical images increases as the α value increases. However, increasing the α value results in a corresponding decrease in the accuracy, which is much more significant for *MMR-balanced* results.

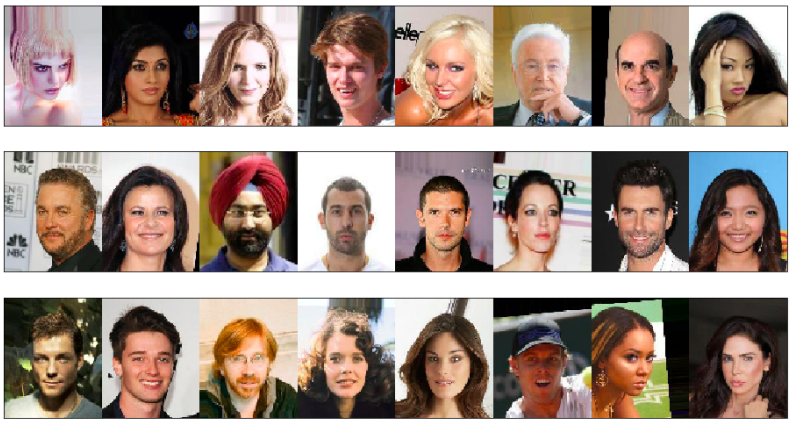
Results for different summary sizes

Once again, we provide the performance of our algorithms and baselines for different summary sizes. For *QS-balanced*, *MMR-balanced* and the baselines, we look at the average fraction of images of gender anti-stereotypical and dark-skinned people in the top k results, where k ranges from 2 to 50; the average is taken over all occupations. The results are presented in Figure A.21. The figures show that *QS-balanced* returns a larger fraction of gender anti-stereotypical images for all summary sizes, compared to all baselines, other than *AUTOLABEL*. While *AUTOLABEL* is able to achieve better gender diversity in this case, due to the good performance of the auto-gender classifier, simply using the partitions has an impact on the accuracy of the summaries generated by *AUTOLABEL*.



(a) Control Set - 1

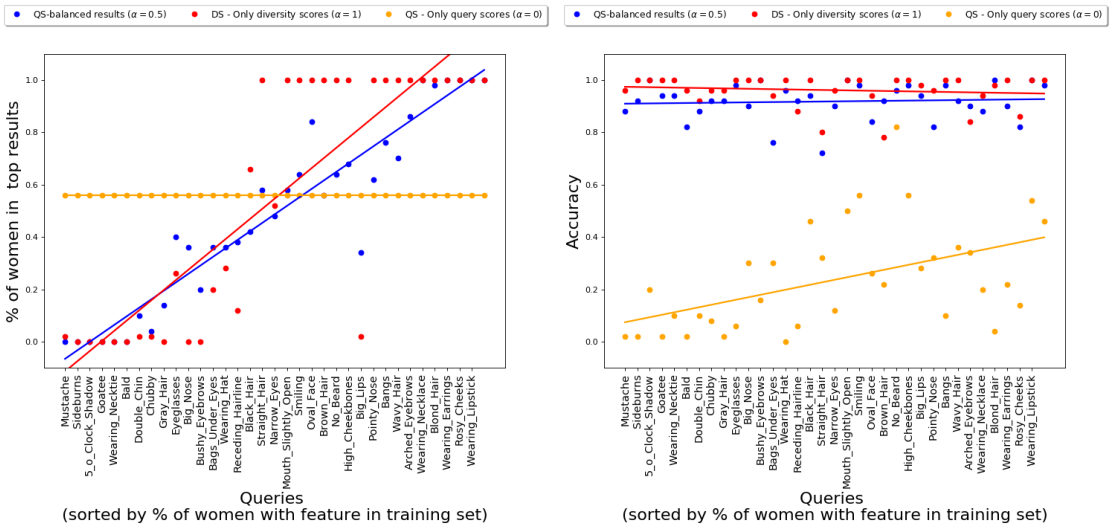
(b) Control Set - 2



(c) Control Set - 3

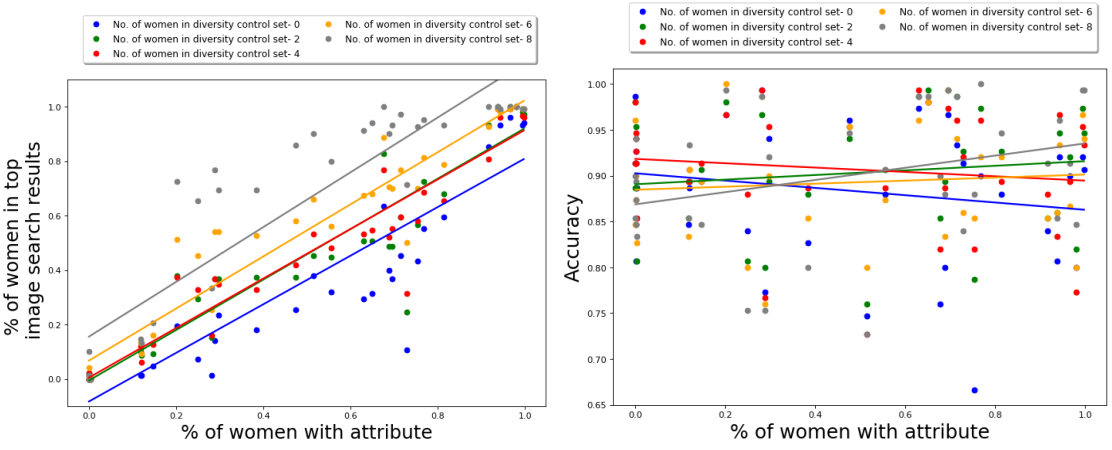
(d) Control Set - 4

Figure A.16: CelebA dataset: Control Sets used in the for empirical evaluation on CelebA dataset.



(a) Gender diversity comparison (b) Accuracy comparison

Figure A.17: CelebA dataset: Gender and accuracy comparison of results of *QS-balanced* algorithm for all queries.



(a) Gender comparison (b) Accuracy comparison

Figure A.18: CelebA dataset: Performance of *QS-balanced* algorithm on control sets with different compositions.

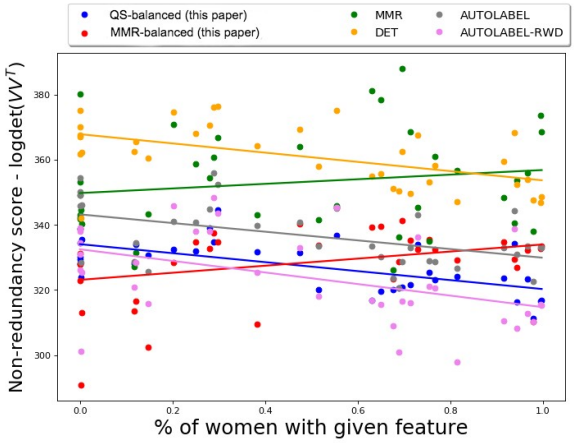
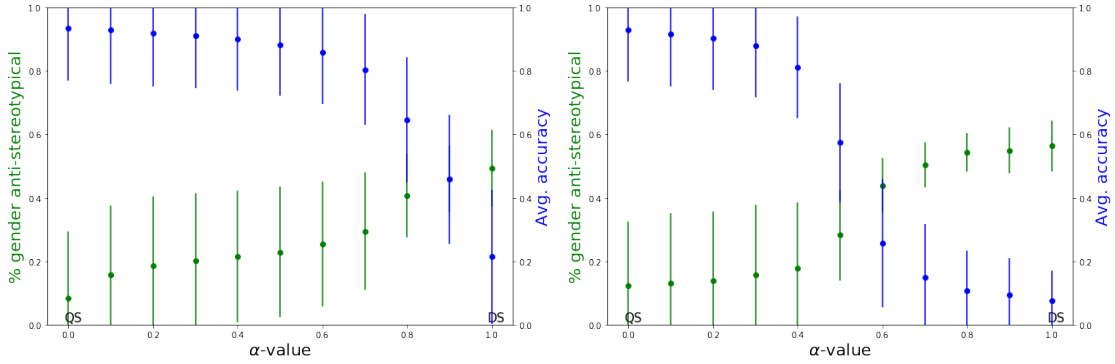
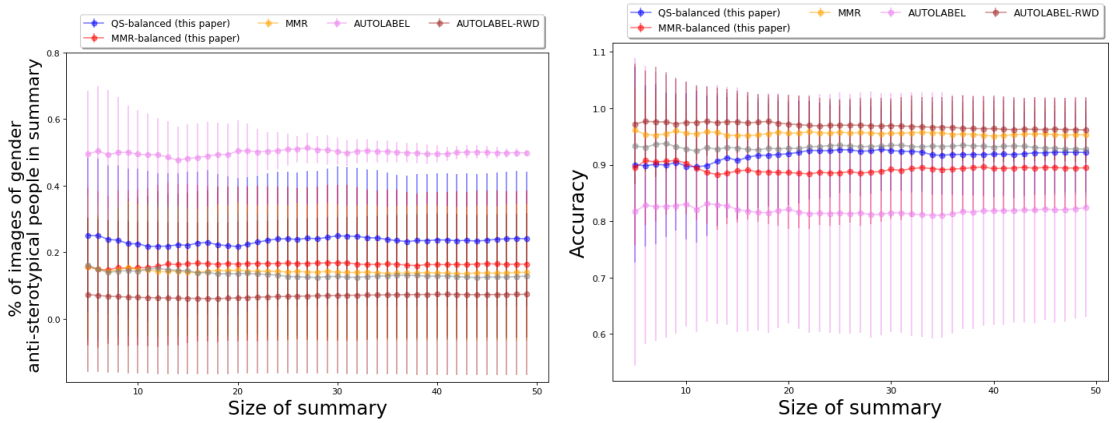


Figure A.19: CelebA dataset: Non-redundancy comparison of our methods vs baselines.



(a) QS-balanced (b) MMR-balanced

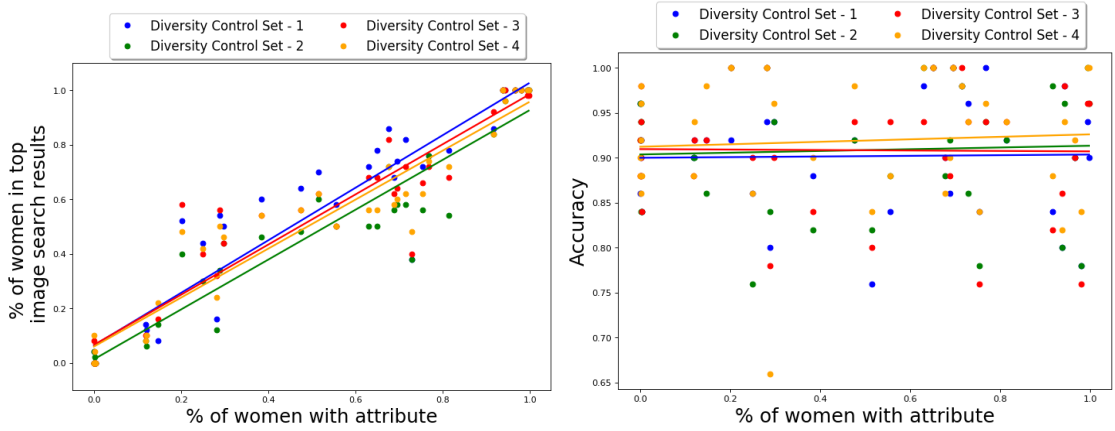
Figure A.20: CelebA dataset: Gender diversity and query similarity comparison of results of *QS-balanced* and *MMR-balanced* algorithms for different α-values.



(a) Fraction of images of gender anti-stereotypical people vs summary size

(b) Accuracy vs summary size

Figure A.21: CelebA dataset: Variation of the fraction of gender anti-stereotypical images and accuracy vs size of summary for all algorithms.



(a) Gender diversity comparison

(b) Accuracy comparison

Figure A.22: CelebA dataset: Gender diversity and accuracy comparison of results of *QS-balanced* algorithm on different control sets. For all the control sets, the performance with respect to gender diversity and accuracy seems to be similar.

A.3 Appendix for Chapter 5

A.3.1 Details of summarization algorithms

TF-IDF. This baseline [203] uses the frequency of the words in a sentence to quantify their weight. However, if a word is very common and occurs in a lot of sentences, then it is likely that the word is part of the grammar structure; hence inverse of document frequency is also taken into account while calculating its score[5]. For any sentence x in S, let $W(x)$ denote the set of words in the sentence. Then the weight assigned to x is $\frac{1}{|W(x)|}\sum_{w\in W(x)} tf(w,x)\cdot \log\frac{|S|}{idf(w,S)}$, where $tf(w,x)$ is the number of times w occurs in x and $idf(w,S)$ is the number of sentences in which w occurs.

Hybrid TF-IDF. The standard TF-IDF has been noted to have poor performance for Twitter posts, primarily due to a lack of generalization of Twitter posts as documents [230]. Correspondingly, a Hybrid TF-IDF [150] approach is proposed that calculates word frequency considering the entire collection as a single document.[5] In other words, the $tf(w,x)$ term in the weight assigned by TF-IDF is replaced by $tf(w,S)$ for Hybrid TF-IDF.

LexRank. This unsupervised summarizer constructs a graph over the dataset, with the similarity between sentences quantifying the edge-weights [104], measured using cosine distance between their TF-IDF word vectors. Using the PageRank algorithm, sentences are then ranked based on how "central" they are within the graph [6].

[5] Internally implemented using the python sklearn and networkx libraries.
[6] `https://github.com/crabcamp/lexrank`

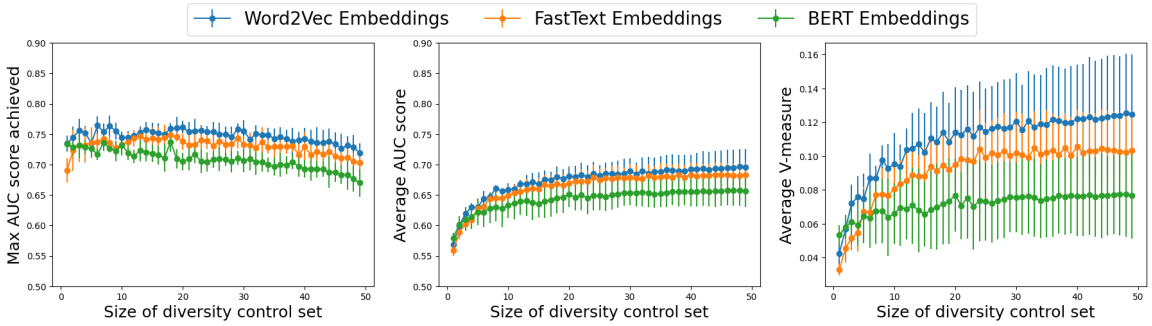
TextRank. This algorithm quantifies the similarity using a modified score of word document frequency [209] and then uses PageRank to rank the sentences; however, it has been shown to achieve better performance for some standard datasets [230][5].

Centroid-Word2Vec. This algorithm assigns importance scores to sentences based on their distance from the *centroid* of the dataset [262] (related to [212]). For vector representation, we use Word2Vec embeddings, pre-trained on a large Twitter dataset [120]. As mentioned in Section 1, it also has a non-redundancy component; if the minimum distance between the feature of the candidate sentence and the feature of a sentence already in the summary is higher than a threshold (0.95 in our case), it is discarded[7].

MMR. This is a post-processing re-ranking algorithm that, at every iteration, greedily chooses the sentence which has the highest MMR score, calculated as the combination of importance score and dissimilarity with the sentences already present in the summary [122, 193]. To get the base importance score, we use the TF-IDF algorithm [5]. Since MMR is a greedy post-processing approach itself, we do not use it as a blackbox algorithm for our framework.

SummaRuNNer. Finally, we use a recent Recurrent Neural Network-based method, SummaRuNNer [224], that treats summarization as a sequential classification problem over the dataset, and generates summaries comparable to the state-of-the-art for the CNN/DailyMail dataset [145]. Since it is not possible to train this model over the Twitter datasets we consider (due to the non-availability of dataset-summary pairs for Twitter dataset), we use the model pre-trained on a standard summarization evaluation dataset [8].

[7] `https://github.com/TextSummarizer/TextSummarizer`
[8] Unofficial implementation: `https://github.com/hpzhao/SummaRuNNer`



(a) Maximum AUC score vs $|T|$ (b) Mean AUC score score vs $|T|$ (c) Maximum V-measure vs $|T|$

Figure A.23: The figure presents how effective different diversity control sets are in clustering posts of the different dialects. Figure (a) presents the average maximum AUC score achieved by a control set across folds for different control set sizes, while Figure (b) presents the mean AUC score achieved by a control set across folds. As an alternative measure, Figure (c) presents the mean V-measure across folds.

Inouye and Kalita [150] empirically analyze the performance of TF-IDF, Hybrid TF-IDF, LexRank, and TextRank on small Twitter datasets (containing only around 1500 tweets for 50 trending topics, not sufficient for a diversity analysis). Their findings suggest that Hybrid TF-IDF produces better summaries for Twitter summarization than TF-IDF, LexRank, and TextRank (as evaluated using ROUGE metrics and manually-generated summaries). For larger and more-recent Twitter datasets, Nguyen et al. [230] found that TextRank and Hybrid TF-IDF have similar performance. Rossiello et al. [262] showed that the centroid-based approach performs better than LexRank, frequency, and RNN-based models on the DUC-2004 dataset. The original papers for most of these algorithms primarily focused on the evaluation of these methods on DUC tasks or CNN/DailyMail datasets; however, the documents in these datasets correspond to news articles from a particular agency and do not usually have significant dialect diversity within them.

Table A.3: Diversity control set for *TwitterAAE evaluations*

AAE tweets
"ATMENTION yea dats more like it b4 I make a trip up der"
"these n***s talmbout money but . really ain't getting no money .. I be laughing at these n***s cause that shit funny ATMENTION"
"Me and Pay got matching coupes, me and kid f***ed ya boo"
"ATMENTION he bites his lips and manages to kick off his remaining clothes"
"Our Dog Is A Big Baby And A Wanna Be Thug EMOJI"
"Its a Damn Shame' iont GangBang but i beat a N*** Blue Black"
"ATMENTION yes, my amazon . Lol Im good . Pop-a-lock came by . Thx!"
"ATMENTION: ATMENTION You talking now? RIGHT? im typing nd texting not talking"
"Soon as u think you gotcha 1 you find out she f***in erbody!!"
"ATMENTION lmaooooooooooooooooooooo, that was the funniest shit ever to hit twitter dawg :D swearrr .. But yall do yall thang"
"Yea Ill Be Good In Bed But Ill Be Bad To Ya!"
"ATMENTION nope tell her get dressed im bouta come get her lol"
"Now omw to get my hair done for coronation tomorrow"
"Ohhhh Hell Naw Dis B**** Shay Got My Last Name * Johnson *"
WHE tweets
"You don't have to keep on smiling that smile that's driving me wild"
"ATMENTION it's probably dead because he hasn't texted me back either"
"ATMENTION amen . Honestly have trouble watching that movie . Just because of her."
"I need to get on a laptop so I can change my tumblr bio"
"Shout out to the blue collar workers . Gotta love it"
"Jax keeps curling up on my bed and tossing and turning repeatedly . Like he cant get comfy . #Soocute #Puppylove"
"ATMENTION you just can't go wrong with Chili's . They serve a mean chips and salsa"
"ATMENTION Tenuta hasn't been good since he left GT and he hates recruiting"
"ATMENTION: Probably the coolest thing I can do ATMENTION yeah, pretty frickin' sweet! Thanks"
"ATMENTION you said we were hanging all day...Lol I don't have a car alslo"
"I want a love like off The Vow .. #perfect #oneday"
"Philosophy is the worst thing to ever happen to the world"
"How come I can never get in a " gunning " fight with anyone? #Jealous"
"'Poor poor Merle, bravo for Michael Rooker and Norman Reedus's performance on last night's show.'

A.3.2 Choice of diversity control set

In this section, we provide a method to construct a *good* diversity control set. For this analysis, we limit ourselves to assessing diversity with respect to AAE and WHE dialects. We employ a smaller processed version of the TwitterAAE dataset, containing 250 AAE posts and 250 WHE (provided by [31]), to select diversity control sets.

Evaluation details. The size of the diversity control set should ideally be much smaller than the evaluation dataset; this will assist in better curation of the control sets. Hence, we restrict the size of the control sets for our simulations to be at most 50.

We perform a 5-fold cross-validation setup for this simulation. For each fold, we have a validation partition U of 400 posts and a train partition of 100 posts (both containing an equal number of AAE and WHE posts); we use the train partition to construct a diversity control set. We sample a set of posts from the train partition, making sure that the set has an equal number of AAE and WHE posts, and use it as a diversity control set; let T denote this set of posts. Then for each $x_c \in T$ and $x \in U$, we calculate the score $\text{sim}(x_c, x)$, and to each $x \in U$, we assign the dialect label of the post $\arg\max_{x_c \in T} \text{sim}(x_c, x)$. Finally, for this prediction task, we report the AUC score and V-measure between the assigned and true dialect labels for posts in U. AUC refers to the area under the Receiver Operating Characteristic (ROC) curve. It is a measure commonly used to evaluate how the performance of a binary learning task. V-measure, on the other hand, is used to evaluate clustering tasks [261]. This measure combines homogeneity (the extent to which AAE clusters contain AAE posts) and completeness (all AAE posts are assigned to AAE clusters). We repeat the sample-and-predict process 50 times for each fold, and we record the max, mean, and standard error of AUC and V-measures across all

repetitions.

To calculate similarity $\text{sim}(z, x)$ between two sentences, we will use the pre-trained word and sentence embeddings to find the feature vectors for these sentences, and then measure the similarity as $1 -$ cosine-distance between the feature vectors. We employ three popular and robust pre-trained embeddings for this task: (a) Word2Vec [210], (b) FastText [34], and (c) BERT embeddings [87]. Using Word2Vec and FastText model, we obtain word representations; to obtain sentence embeddings from word representations, we use the aggregation method of Arora et al. [15] which computes the weighted average of the embeddings of the words in the sentence, where the weight assigned to a word is proportional to the smooth inverse frequency of the word. For Word2Vec and FastText, we use the models pre-trained on a corpus of 400 million posts [120]. Output from the second-last hidden layer of the pre-trained BERT model can be used to directly obtain sentence embeddings.

Results. Figure A.23 shows that diversity control sets constructed in this manner are indeed suitable for differentiating between posts of different dialects. Plot A.23a shows that *good control sets* are able to achieve AUC scores greater than 0.8 (including the one presented in Table A.3). Furthermore, the average AUC score is also greater than 0.65 for diversity control set sizes greater than 10, implying that diversity control sets of sizes between 10 and 50 are indeed suitable for this task. Given that the diversity control sets do perform fairly well on this clustering task, this provides further insight into the improved dialect diversity when using our post-processing framework with standard summarization algorithms as blackbox. Secondly, Word2Vec embeddings achieve better performance than FastText and BERT embeddings and, hence, we use Word2vec representations for the empirical analysis of our framework as well.

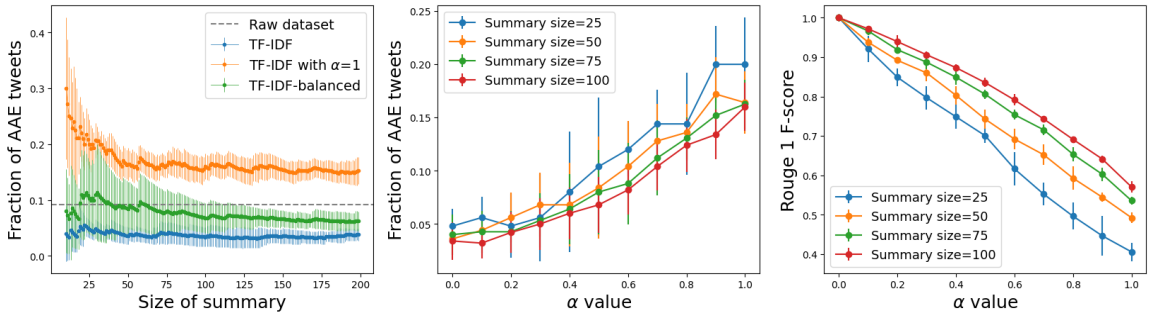
Using the above method, we construct a diversity control set of size 28 for *TwitterAAE Evaluations* (Table A.3), a control set of size 40 for *Crowdflower Evaluations* (Table A.4), and control set of size 20 for *Claritin Evaluations* (Table A.5).

A.3.3 Other details and results for TwitterAAE dataset

The control set used for TwitterAAE simulations is provided in Table A.3.

Evaluation of our model on random collections of TwitterAAE datasets

For random collections of the TwitterAAE dataset, with different fractions of AAE tweets in them, we use our model to generate summaries of different sizes. The results for TF-IDF are given in Figure A.24 and A.25; for Hybrid-TF-IDF, see Figure A.26 and A.27; for LexRank, see Figure A.28 and A.29; for TextRank, see Figure A.30 and A.31; for SummaRuNNer, see Figure A.32 and A.33. $\alpha = 0.5$, unless mentioned otherwise.

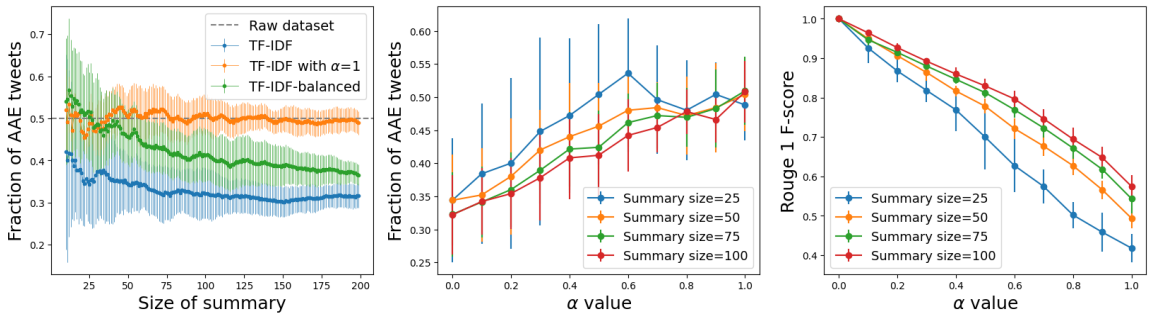


(a) AAE frac. vs summary size (b) AAE frac. in summary vs α (c) Rouge-1 F-score vs α

Figure A.24: Evaluation of our model on datasets containing 8.7% AAE tweets using TF-IDF as algorithm A.

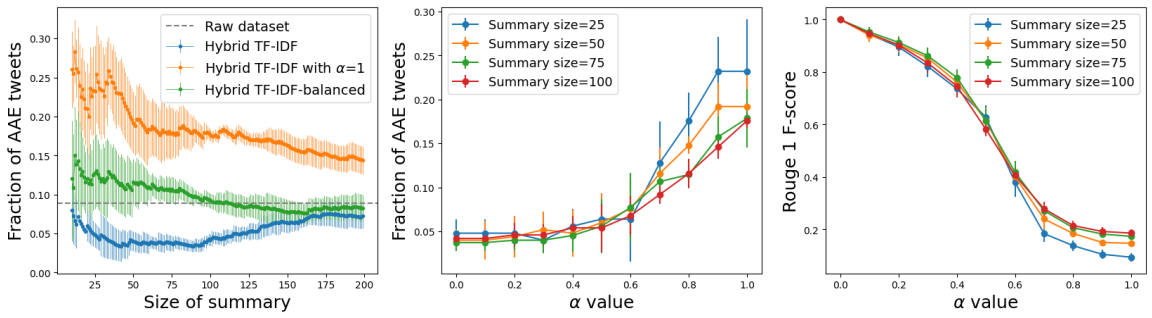
Evaluation of our model on keyword-specific collections of TwitterAAE

Next, we also present the results for our model on collections of TwitterAAE dataset containing the keywords used in Section 5.2. The results for TF-IDF are given in



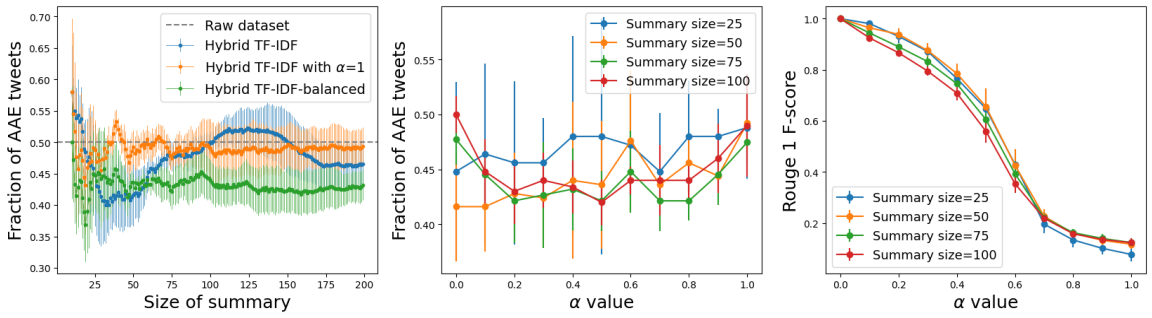
(a) AAE frac. vs summary size (b) AAE frac. in summary vs α (c) Rouge-1 F-score vs α

Figure A.25: Evaluation of our model on datasets containing 50% AAE tweets using TF-IDF as algorithm A.



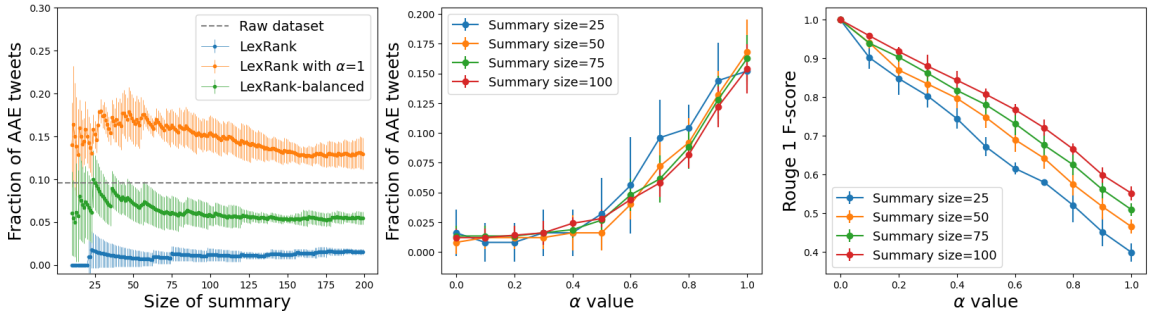
(a) AAE frac. vs summary size (b) AAE frac. in summary vs α (c) Rouge-1 F-score vs α

Figure A.26: Evaluation of our model on datasets containing 8.7% AAE tweets using Hybrid TF-IDF as algorithm A. Here $\alpha = 0.7$ for balanced algorithm



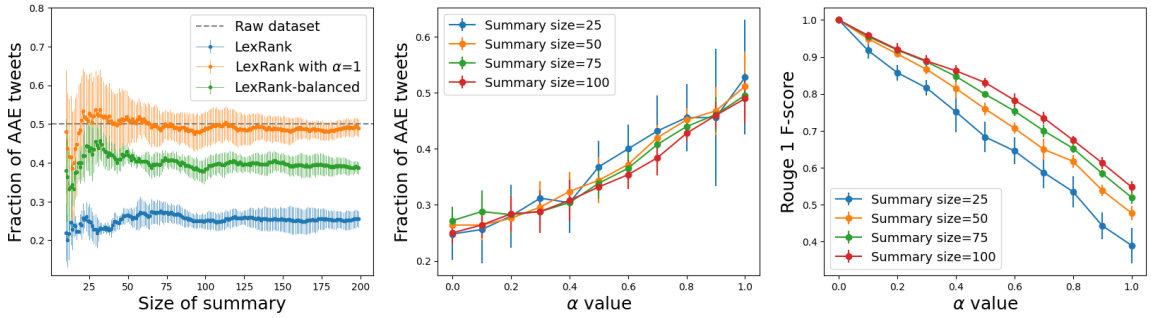
(a) AAE frac. vs summary size (b) AAE frac. in summary vs α (c) Rouge-1 F-score vs α

Figure A.27: Evaluation of our model on datasets containing 50% AAE tweets using Hybrid TF-IDF as algorithm A. Here $\alpha = 0.7$ for balanced algorithm.



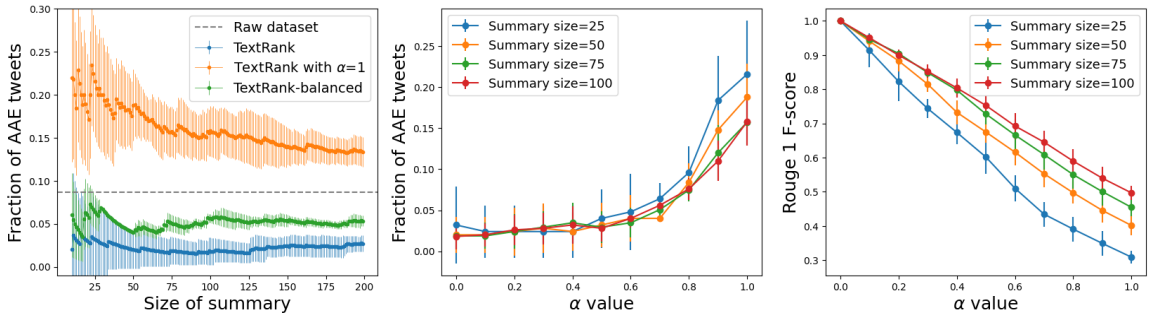
(a) AAE frac. vs summary size (b) AAE frac. in summary vs α (c) Rouge-1 F-score vs α

Figure A.28: Evaluation of our model on datasets containing 8.7% AAE tweets using LexRank as algorithm A. Here $\alpha = 0.7$ for balanced algorithm.



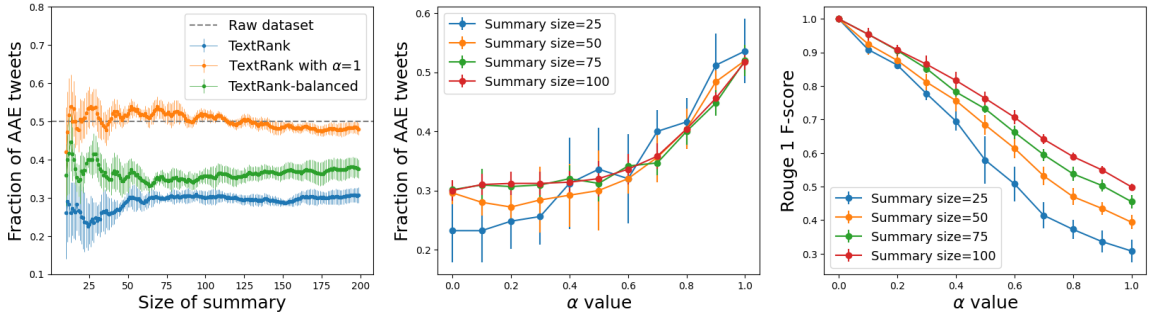
(a) AAE frac. vs summary size (b) AAE frac. in summary vs α (c) Rouge-1 F-score vs α

Figure A.29: Evaluation of our model on datasets containing 50% AAE tweets using LexRank as algorithm A. Here $\alpha = 0.7$ for balanced algorithm.



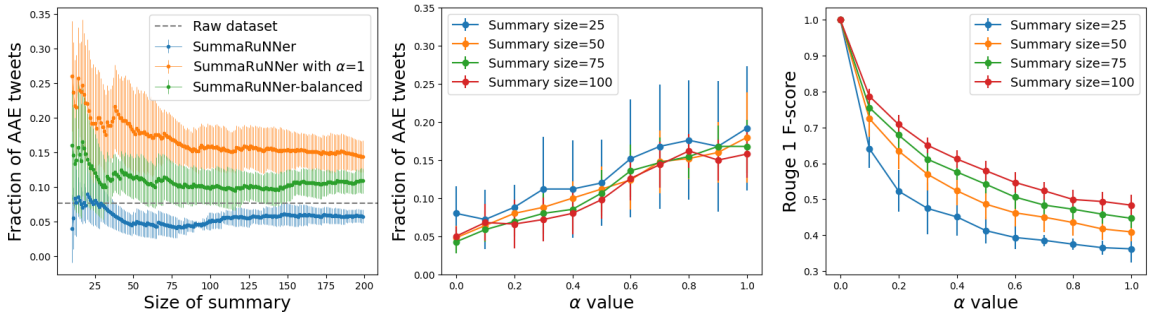
(a) AAE frac. vs summary size (b) AAE frac. in summary vs α (c) Rouge-1 F-score vs α

Figure A.30: Evaluation of our model on datasets containing 8.7% AAE tweets using TextRank as algorithm A. Here $\alpha = 0.7$ for balanced algorithm.



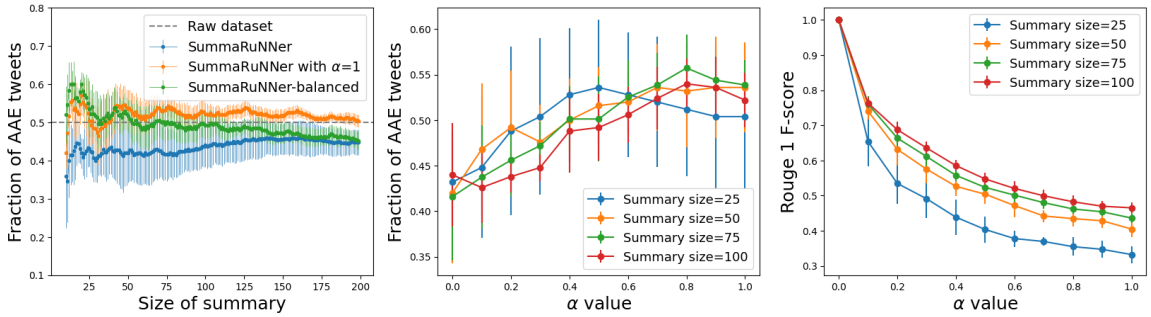
(a) AAE frac. vs summary size (b) AAE frac. in summary vs α (c) Rouge-1 F-score vs α

Figure A.31: Evaluation of our model on datasets containing 50% AAE tweets using TextRank as algorithm A. Here $\alpha = 0.7$ for balanced algorithm.



(a) AAE frac. vs summary size (b) AAE frac. in summary vs α (c) Rouge-1 F-score vs α

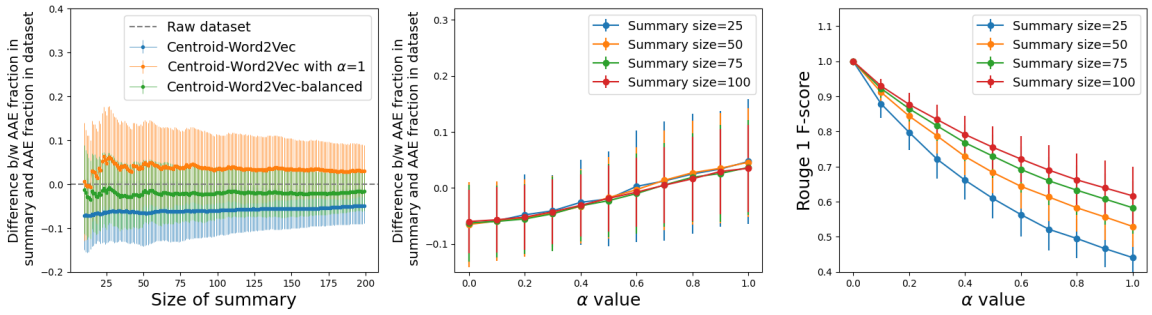
Figure A.32: Evaluation of our model on datasets containing 8.7% AAE tweets using SummaRuNNer as algorithm A.



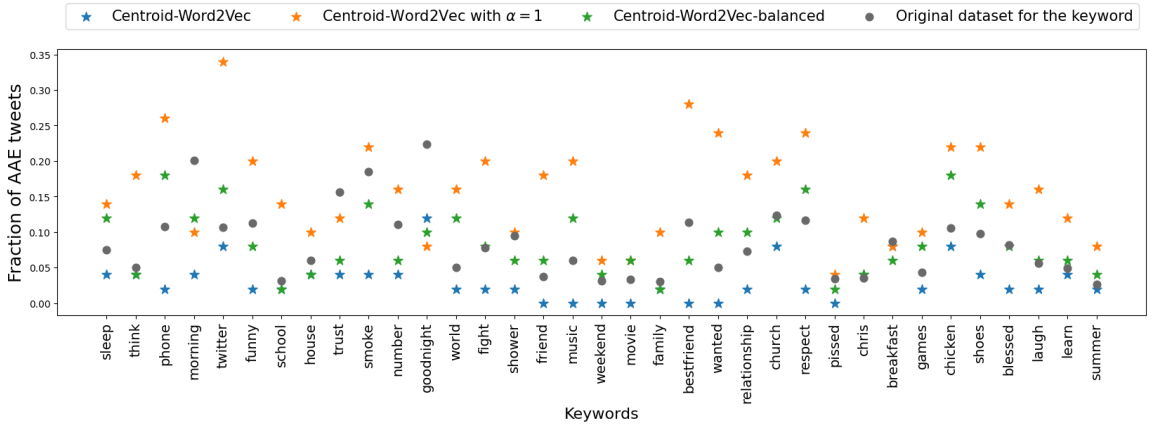
(a) AAE frac. vs summary size (b) AAE frac. in summary vs α (c) Rouge-1 F-score vs α

Figure A.33: Evaluation of our model on datasets containing 50% AAE tweets using SummaRuNNer as algorithm A.

Figure A.35; for Hybrid-TF-IDF, see Figure A.36; for LexRank, see Figure A.37; for TextRank, see Figure A.38; for Centroid-Word2Vec, see Figure A.34; for SummaRuNNer, see Figure A.39.



(a) AAE frac. vs summary size (b) AAE frac. in summary vs α (c) Rouge-1 F-score vs α

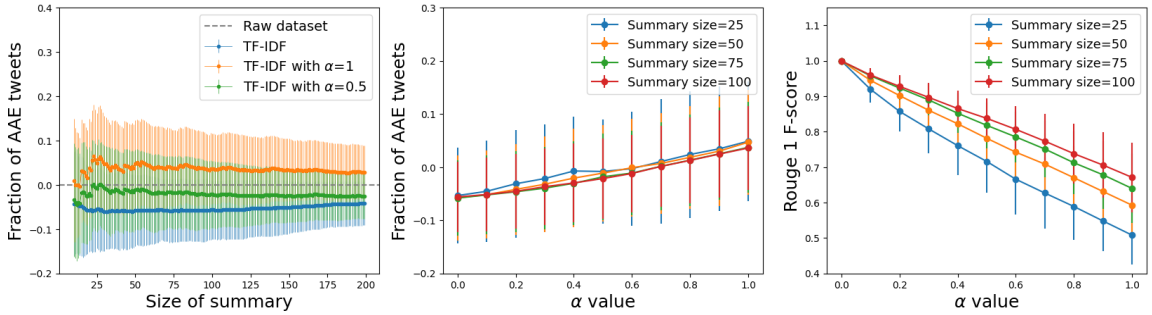


(d) AAE fraction for different keywords and summary size = 50

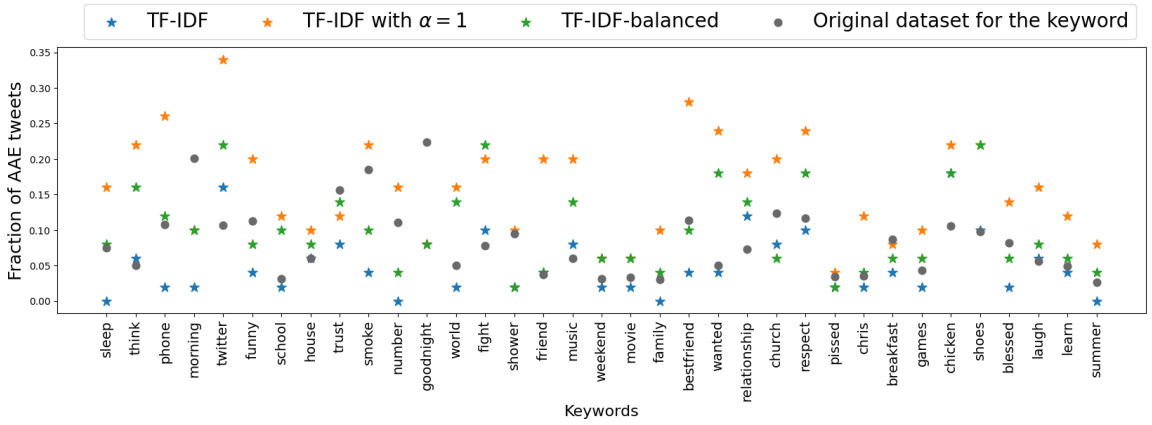
Figure A.34: Evaluation of our model on keyword-specific datasets using Centroid-Word2Vec as A.

Evaluation of our model using different diversity set compositions

We also present the evaluation for the setting where the diversity control set has an unequal fraction of AAE and WHE posts. For random collections where the fraction of AAE posts in the collection is 50%, Figure A.40. As expected, the fraction of AAE posts in summary increases as the fraction of AAE posts in the control set increases. This is another parameter that can be tuned to adjust and obtain the

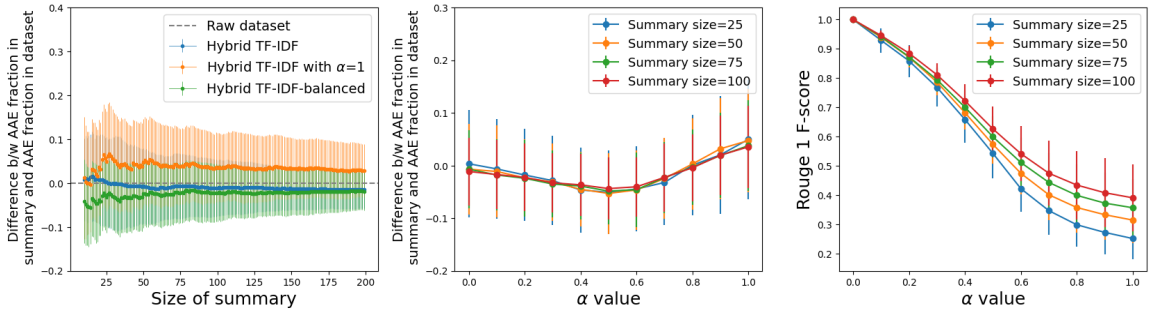


(a) AAE frac. vs summary size (b) AAE frac. in summary vs α (c) Rouge-1 F-score vs α

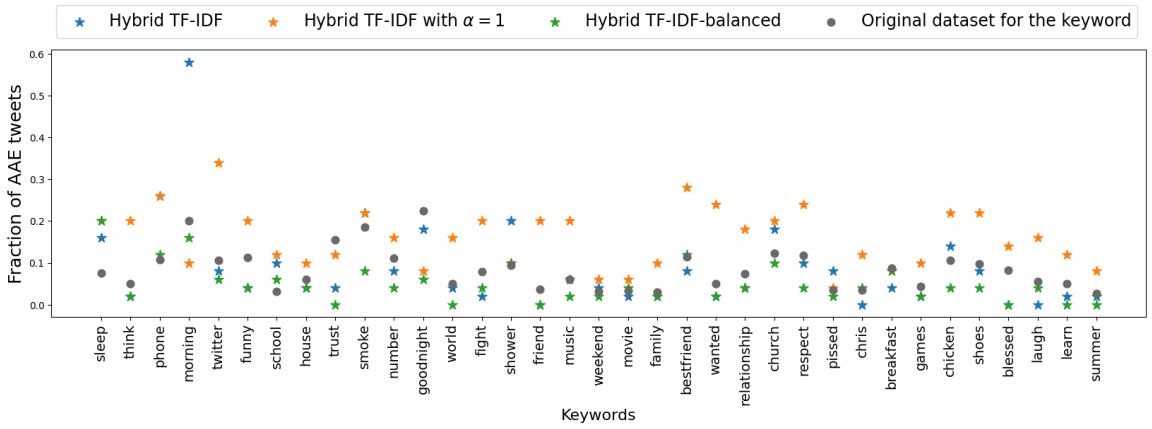


(d) AAE fraction for different keywords and summary size = 50

Figure A.35: Evaluation of our model on keyword-specific datasets using TF-IDF as A.

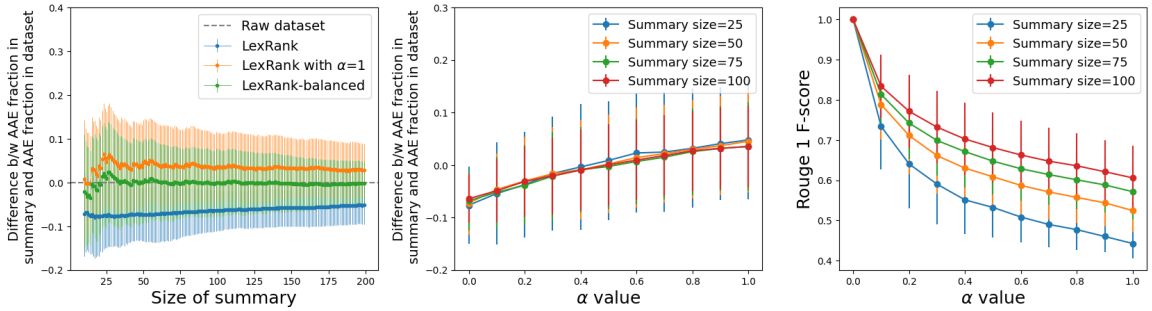


(a) AAE frac. vs summary size (b) AAE frac. in summary vs α (c) Rouge-1 F-score vs α

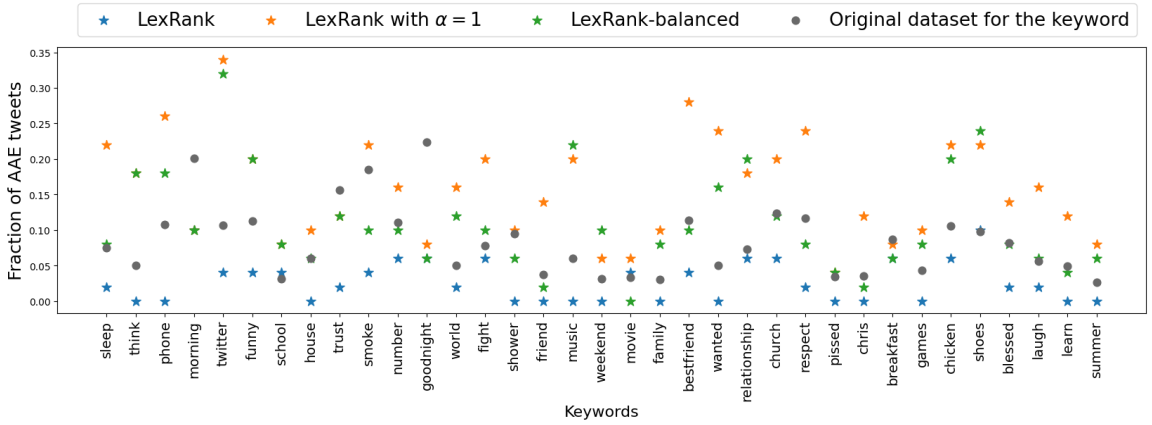


(d) AAE fraction for different keywords and summary size = 50

Figure A.36: Evaluation of our model on keyword-specific datasets using Hybrid TF-IDF as algorithm A. Here $\alpha = 0.7$.

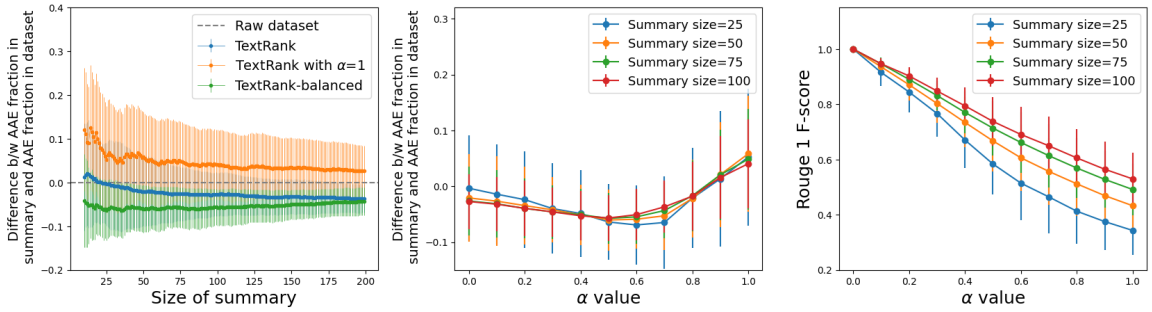


(a) AAE frac. vs summary size (b) AAE frac. in summary vs α (c) Rouge-1 F-score vs α

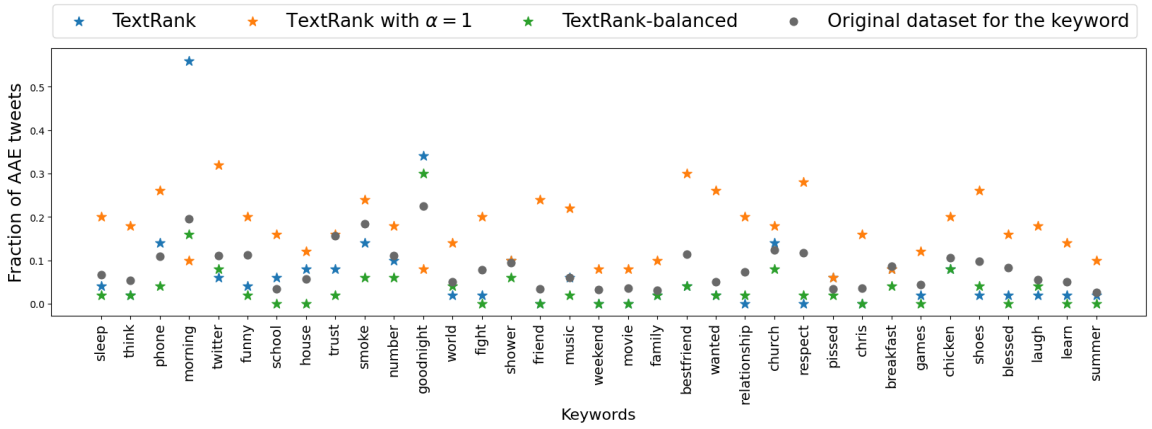


(d) AAE fraction for different keywords and summary size = 50

Figure A.37: Evaluation of our model on keyword-specific datasets using LexRank as A.

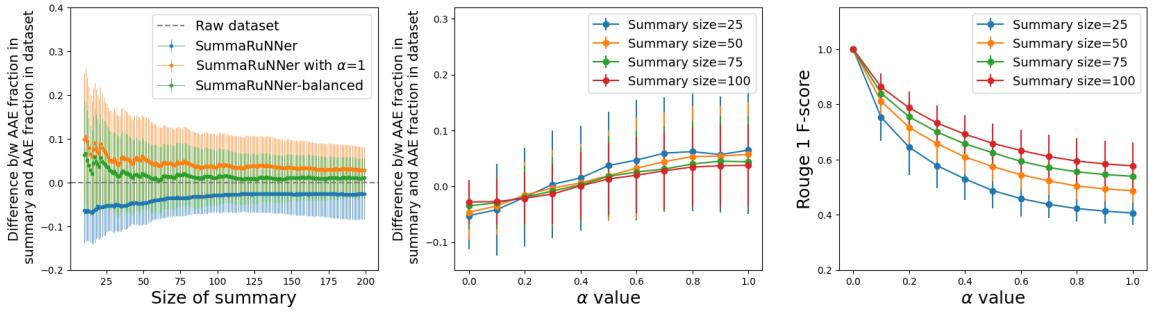


(a) AAE frac. vs summary size (b) AAE frac. in summary vs α (c) Rouge-1 F-score vs α

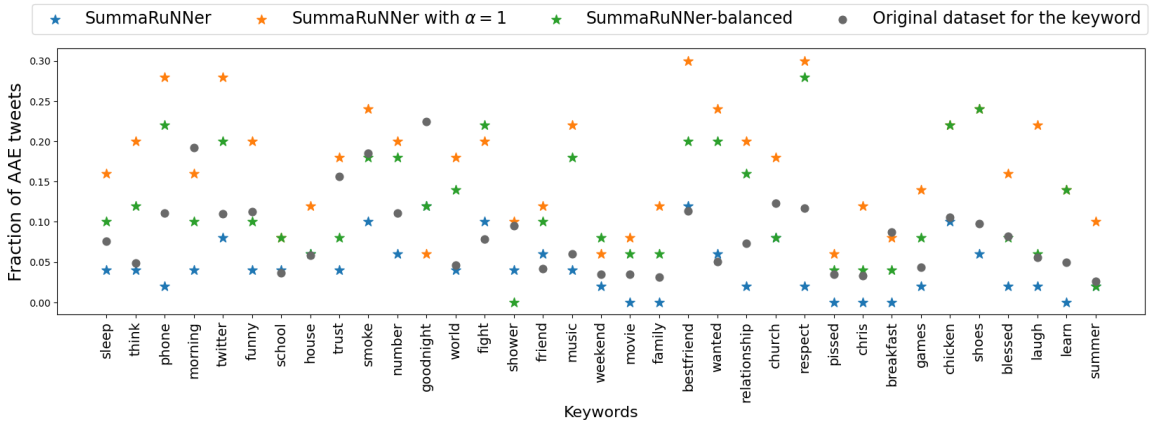


(d) AAE fraction for different keywords and summary size = 50

Figure A.38: Evaluation of our model on keyword-specific datasets using TextRank as algorithm A. Here $\alpha = 0.7$ for balanced algorithm.



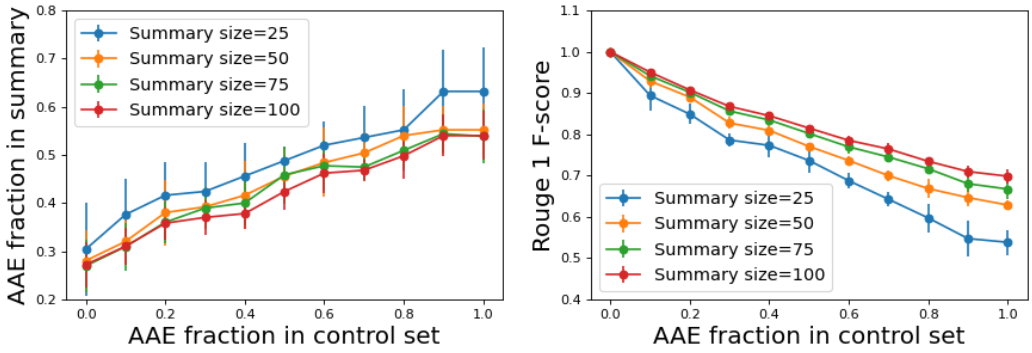
(a) AAE frac. vs summary size (b) AAE frac. in summary vs α (c) Rouge-1 F-score vs α



(d) AAE fraction for different keywords and summary size = 50

Figure A.39: Evaluation on keyword-specific datasets using SummaRuNNer as A.

desired fraction of AAE posts in the summary.



(a) AAE fraction in summary vs control set

(b) Rouge-1 F-score vs control set

Figure A.40: Evaluation of our model using different control set compositions.

A.3.4 Other details and results for Crowdflower Gender AI dataset

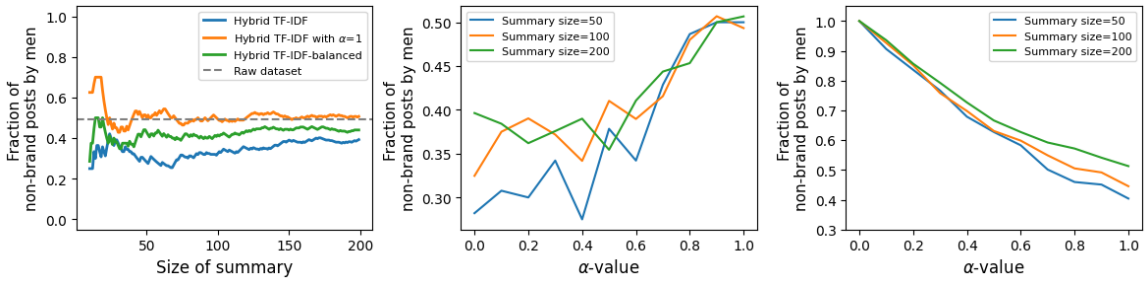
The diversity control set used for Crowdflower Gender evaluation is presented in Table A.4.

Evaluation of our model with different blackbox algorithms

The performance of our model using different blackbox algorithms is presented here. The results for Hybrid TF-IDF are given in Figure A.41; for LexRank, see Figure A.42; for TextRank, see Figure A.43; for Centroid-Word2Vec, see Figure A.44.

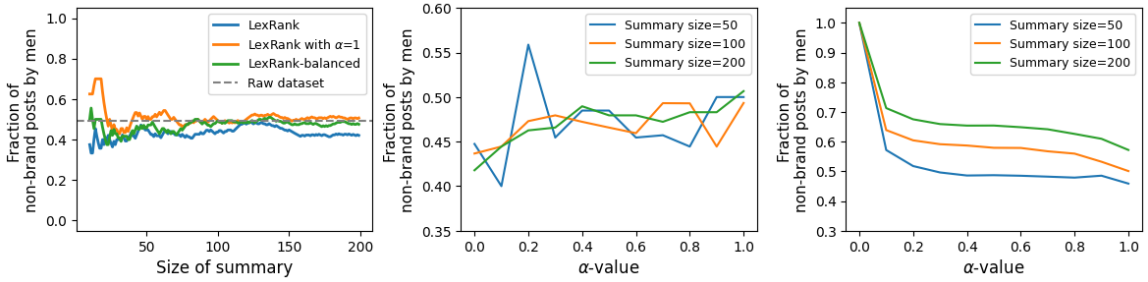
Table A.4: Diversity control set for simulations on Crowdflower Gender AI dataset

Tweets by female user-accounts
“jameslykins haha man! the struggle is reeeeeal! ”
“red lips and rosy cheeks”
“#mood spirit of jezebel control revelation 21820, 26 a war goes on in todays church, and the ”
“where the hell did october go? halloween is already this weekend. ”
“my lipstick looked like shit and my hair is usually a mess but im still cute tho so ”
“say she gon ride for me , ill buy the tires for you ”
“so excited to start the islam section in my religions class ”
“wow blessed my 200 kate spade bag is ripping and ive only used it twice a week since the end of september .”
“all ive done today is lie around and homework tbh”
“of course you want to blame me for not finishing college and thus bringing this debt to myself of course”
“misskchrista everyone was obsessed with rhys though, no one really knew the other two xxx”
“papisaysyes at first i thought this said, my d**k is on drugs and i still dont know which is worse lol”
“huge announcement and #career change for 2016. #goals #dreams #nymakeupartist ”
“practice random acts of kindness and make it a habit #aldubpredictions”
“ sammanthae glad i can make you laugh i miss you and love you too!!”
“nba i play basketball to escape reality. between the exercise and the diff personalities memories are made!”
“z100newyork please let me attend the future now vip party tonight i love demi and nick #z100futurenow ”
“#win 2 random jumbies stuffed animals #giveaway us only 1113 bassgiraffe ”
“daynachirps thats a great point. thanks for the reminder. #contentchat”
“ive told bri all this time it would happen and it finally did ”
Tweets by male user accounts
“warrenm ill be using my new mbp. i do see dells 5k line needs 2 thunderbolt connections to make it a true 5k display. not the case here?”
“logic301 salute on the new visuals my g! dope as f**k”
“i liked a youtube video official somewhere over the rainbow 2011 israel iz kamakawiwoole”
“laughs and cries at the same time cause true ”
“akeboshi night and day”
“now you all know the monster mash, but now for something really scary, the climate mash ”
“i hate when u tell someone u love them and they ignore u ”
“the finger hahsah ”
“the corruption of the wash. d.c. crowd is now of epic proportions. enlist gt join us ”
“i wish i went to school closer to mark a schwab . beating up doors and walls looks like a lot of fun.”
“keepherwarm kobrakiddlng aimhbread now ill let you know that ive known a guy my whole life who dated several girls and then later on”
“xavierleon fr like wtf are they taking that they just cant f***ing dye and busting through doors?! ”
“heh, i just remember people actually think that se and hp are intentionally sabotaging the football team.”
“we must lessen the auditory deprivation! i agree earlier the implantation, the better! ”
“#repost seekthetruth with repostapp. repost ugly by nature 85 of the #tampons, cotton and ”
“the #ceo needs to embrace and sell social to the team or else is goes nowhere. bernieborges #h2hchat #ibminsight ”
“if you scored a touchdown on sunday and didnt dab, hit them folks, or do that hotline bling dance, it shouldnt have counted.”
“zbierband yo zbb, played our last seasonal gig at st. jude. good times had by all. remember the more you drink, the better we sound!”
“i hate writing on the first page of a notebook i feel like im ruining something so perfect”
“we schools should be given credit for growth in the apr, but growth is not the destination. michael jones moboe. ”



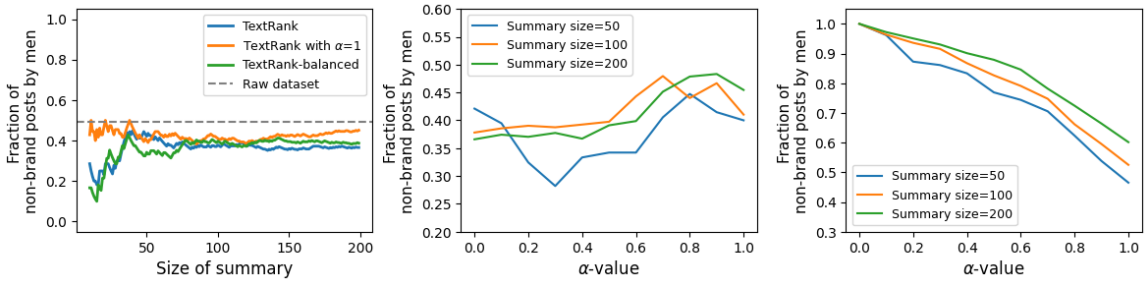
(a) Gender frac. v summary size (b) Gender fraction vs α (c) Rouge-1 F-score vs α

Figure A.41: Evaluation of our model on Crowdflower Gender AI dataset using Hybrid TF-IDF as algorithm A.



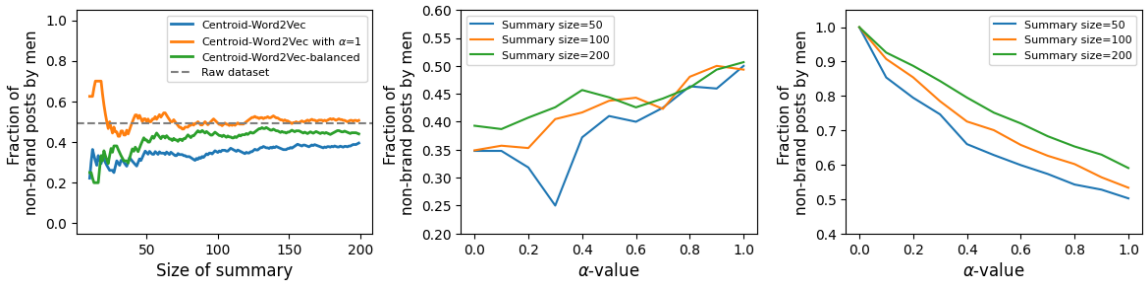
(a) Gender frac. v summary size (b) Gender fraction vs α (c) Rouge-1 F-score vs α

Figure A.42: Evaluation of our model on Crowdflower Gender AI dataset using LexRank as algorithm A.



(a) Gender frac. v summary size (b) Gender fraction vs α (c) Rouge-1 F-score vs α

Figure A.43: Evaluation of our model on Crowdflower Gender AI dataset using TextRank as algorithm A.



(a) Gender frac. v summary size (b) Gender fraction vs α (c) Rouge-1 F-score vs α

Figure A.44: Evaluation of our model on Crowdflower Gender AI dataset using Centroid-Word2Vec as algorithm A.

A.3.5 Diversity control set used for Claritin dataset

The diversity control set used for Claritin Gender evaluation is presented in Table A.5.

Table A.5: Diversity control set for simulations on Claritin dataset

Tweets by female user-accounts
"claritin, why didnt you work? i was desperate thats why i took you. ang mahal mo pa man din! "
"ATMENTION been there. always. done that. youll be fine. claritin works for that. "
"ATMENTION all allergy meds raise als blood pressure a lot. claritin isnt so bad but still sucks. the kid stuff is half dose and works "
"k time to bust out that claritin. siiigh "
"ATMENTION if they are asking for allegra, mucinex, or claritin. they want the d. ATMENTION "
"if a girl sends you a text, heyy, im sick. . she probably wants the d claritin d #pervs "
"ATMENTION yes, claritin, tylenol and ibuprofen. "
"what ever happened to jeff corwin? supposedly he does claritin commercials now. "
"deffo allergic to tingle creams now not on my legs back or belly though but on my arms chest ampface need to buy claritin amp chamomile lotion "
"ATMENTION awesome i never wear glasses so this has sucked doc said taking one claritin dried up my tears. just one?? "
Tweets by male user accounts
"ATMENTION i have one xd if she has allergies.. give here some claritind ! "
"if a girl tells you shes sick she wants the d, claritind ATMENTION "
"ATMENTION givin complementary claritin d pills amp shit. "
"claritin and food please #sniffle "
"ok so 2 pills of allegra is not helping my allergies, anyone have another pill i should try? claritin is out "
"she feeling sick? she wants the d. claritind "
"yeah my allergies are acting up , i didnt take any claritin today ATMENTION "
"ATMENTION if a girl sends you hey, im sick. she probably wants the claritind. haha. "
"clearly. claritin clear "
"her allergies were acting up, so i gave her the d.... claritin d. "

A.4 Appendix for Chapter 6

A.4.1 Details of baselines

LL Algorithm [190]

This algorithm, proposed by Li and Liu [190], takes as input a single measure of reliability for each expert and returns k experts using a formula that takes into account the reliability, the number of classes, and size of desired committee size k (see Algorithm 1 in [190]). To calculate the measure of reliability for each expert using the training set, we simply calculate the accuracy of each expert over the training set.

Note that the main drawbacks of this approach are that it simply returns a single committee, i.e. does not choose the experts in an input-specific manner and that it treats the pre-trained classifier as yet another learner.

CrowdSelect Algorithm [249]

CrowdSelect is a more advanced task-allocation algorithm that takes into account the error models of different experts, as well as, their task-specific reliabilities and the individual costs associated with each expert consultation. However, the proposed algorithm assumes that *error rate of workers* for any given task is provided as input or can be estimated using autoregressive methods that use the task identities. In our setting, the specific task classification (for example, cluster identity in case of Section 6.3.1)) may not be available; hence, these error models need to be separately constructed.

To construct the error models for the experts, for each expert i, we simply train a two-layer neural network h_i on the train feature vectors using binary class labels that correspond to whether the expert's prediction for the given train feature was correct or not. Then, for any test/future sample, h_i will return the probability

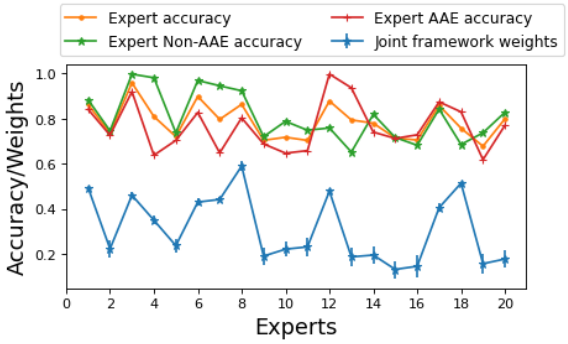


Figure A.45: Weights assigned by the joint learning model and the accuracies of the 20 experts (one iteration shown). Accuracies and weights are seen to follow a similar pattern.

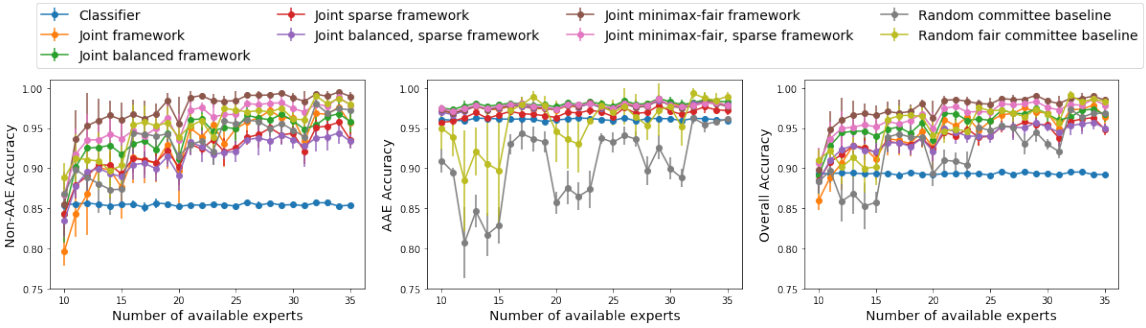


Figure A.46: Performance of all methods for different number of available experts.

that the expert i returns a correct prediction. Using these error models, we then implement Algorithm 1 in [249] to get input-specific committees.

There are three drawbacks to this approach: (1) the pre-trained classifier is once again treated as yet another learner, (2) it is only applicable for binary classification ([249] propose studying extensions to non-binary as future work), and (3) the error models of all experts are learned independently - this is inhibitory since it does not allow the perfect stratification of input domain into the domains of different experts.

Our method addresses all three drawbacks by learning a single deferrer and learning it simultaneously with a classifier.

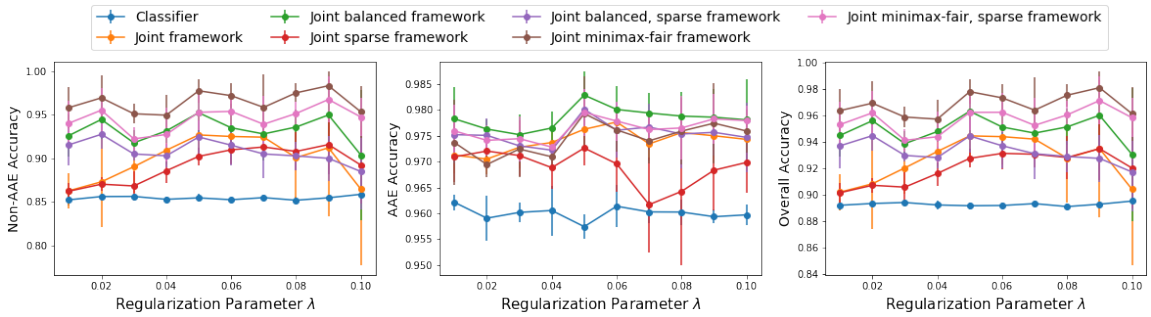


Figure A.47: Performance of all methods for different values of regularization parameter λ.

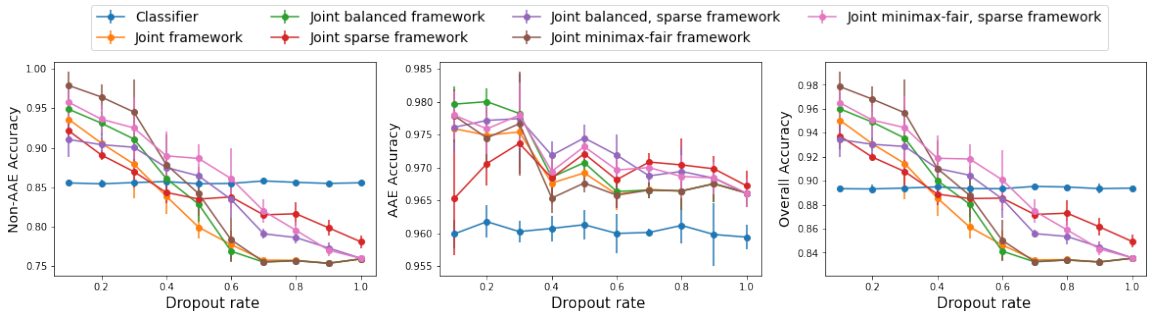


Figure A.48: Performance of all methods for different dropout rates.

A.4.2 Other empirical results for offensive language dataset with synthetic experts

In this section, we present additional empirical results for the offensive language dataset with multiple synthetic experts.

Variation with number of experts

We vary the number of experts m from 10 to 35, while keeping λ fixed at 0 and the dropout rate fixed at 0.2, and present the variation of overall and dialect-specific accuracies when using a different number of experts. The other parameters are kept to be the same as Section 6.3. The results are presented in **Figure A.46**. As expected, the performance of all joint frameworks increases with an increasing number of experts, and the performance of the minimax-fair framework is better

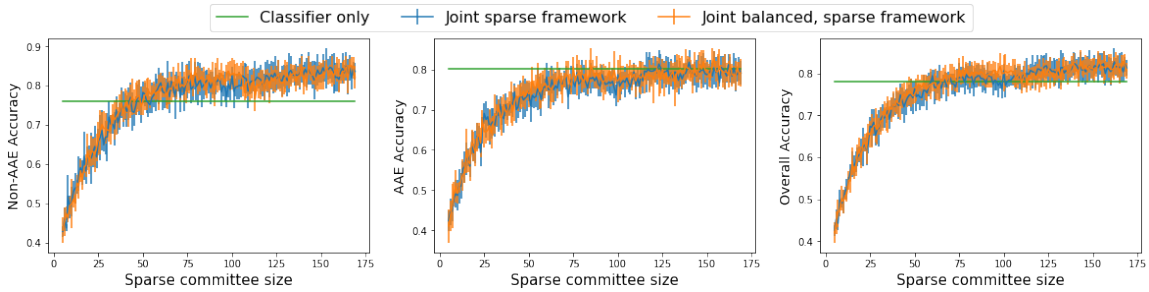


Figure A.49: Performance of sparse variants of the joint frameworks on the MTurk dataset for different committee sizes k.

than other methods in most cases.

Performance of random committee baselines

Figure A.46 also provides further insight into the random committee baselines. Since 75% of the experts are biased against the AAE dialect, simply choosing the committee randomly leads to reduced accuracy for the AAE dialect. When the committee is selected in a dialect-specific manner (random fair committee baseline), the disparity across dialects reduce but the accuracies of the experts are not taken into account. The performance of these two baselines highlights the importance of selecting the experts in an input-specific manner and taking the accuracies/biases of experts while deferring.

Impact of λ

We next vary the parameter λ from 0.01 to 0.1, while keeping the number of experts at $m = 20$ and the dropout rate at 0.2, and present the variation of overall and dialect-specific accuracies for different λ values. The results are presented in **Figure A.47**. The variation with respect to λ shows that setting its value close to 0.05 leads to the best performance for most methods. Smaller values of λ will lead to low dependence on the classifier, while higher values of λ imply associating larger regularization costs with the experts, and the figure shows that the performance

for large λ has larger variance and/or is closer to the performance of the classifier.

Impact of dropout rate

Finally, we vary the dropout rate from 0.1 to 0.9, while keeping the number of experts at $m = 20$ and $\lambda = 0.05$, and present the variation of overall and dialect-specific accuracies for different dropout rates. As expected, larger values of dropout can imply that the framework is unable to decipher the accuracies of the experts and, hence, leads to a drop in accuracy. Reasonable levels of dropout rate (around 0.2), on the other hand, do not impact accuracy but significantly reduce the load on the more accurate experts.

A.4.3 Other empirical results for MTurk dataset

As mentioned in Section 6.4, the task of differentiating between the experts is more challenging for the MTurk dataset since relatively fewer prior predictions are available for each expert. Correspondingly, the sparse variants do not perform so well when the chosen committee size k is small.

The performance of the sparse variants, as a function of k, is presented in **Figure A.49**. From the figure, one can see that to achieve performance similar to or better than the classifier, k needs to be around 60 or larger.