CARACTERIZANDO INTERCONEXÕES E PADRÕES LINGUÍSTICOS NO TWITTER

JOHNNATAN MESSIAS PEIXOTO AFONSO

CARACTERIZANDO INTERCONEXÕES E PADRÕES LINGUÍSTICOS NO TWITTER

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

ORIENTADOR: FABRÍCIO BENEVENUTO DE SOUZA

Belo Horizonte

Junho de 2017

JOHNNATAN MESSIAS PEIXOTO AFONSO

CHARACTERIZING INTERCONNECTIONS AND LINGUISTIC PATTERNS IN TWITTER

Dissertation presented to the Graduate Program in Computer Science of the Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Fabrício Benevenuto de Souza

Belo Horizonte

June 2017

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Characterizing Interconnections and Linguistic Patterns in Twitter

JOHNNATAN MESSIAS PEIXOTO AFONSO

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. FABRÍCIO BENEVENUTO DE SOUZA - Orientador
Departamento de Ciência da Computação - UFMG

PROF. FÁBRÍCIO MURAI FERREIRA
Departamento de Ciência da Computação - UFMG

PROFA. MARIA DA GRAÇA CAMPOS PIMENTEL
Departamento de Ciência da Computação - USP

PROF. WAGNER MEIRA JÚNIOR
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 09 de junho de 2017.

To my future children.

Acknowledgments

I would like to thank my advisor Prof. Dr. Fabrício Benevenuto from Universidade Federal de Minas Gerais (UFMG) who gave me the directions to accomplish my master course. I, also, would like to thank Prof. Dr. Krishma Gummadi who hosted me at Max Planck Institute for Software Systems (MPI-SWS). More importantly, I would like to thank my mother Varlene, father Juarez, my brothers Joarlens and Jeanderson, and my girlfriend Aline for all their support.

Abstract

Social media is considered a democratic space in which people connect and interact with each other regardless of their gender, race, or any other demographic aspect. Despite numerous efforts that explore demographic aspects in social media, it is still unclear whether social media perpetuates old inequalities from the offline world. In this dissertation, we attempt to identify gender and race of Twitter users located in the United States using advanced image processing algorithms from Face++. We investigate how different demographic groups (i.e. male/female, asian/black/white) connect with each other and differentiate them regarding linguistic styles and also their interests. We quantify to what extent one group follows and interacts with each other and the extent to which these connections and interactions reflect in inequalities in Twitter. We also extract linguistic features from six categories (affective attributes, cognitive attributes, lexical density and awareness, temporal references, social and personal concerns, and interpersonal focus) in order to identify the similarities and the differences in the messages they share in Twitter. Furthermore, we extract the absolute ranking difference of top phrases between demographic groups. As a dimension of diversity, we also use the topics of interest that we retrieve from each user. Our analysis shows that users identified as white and male tend to attain higher positions, in terms of the number of followers and number of times in another user's lists, in Twitter. There are clear differences in the way of writing across different demographic groups in both gender and race domains as well as in the topic of interest. We hope our effort can stimulate the development of new theories of demographic information in the online space. Finally, we developed a Web-based system that leverages the demographic aspects of users to provide transparency to the Twitter trending topics system.

Keywords: Twitter, Demographic Aspects, Inequality, Linguistic Patterns, Gender and Race.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Nowadays, millions of people constantly use online social networking sites, such as Facebook and Twitter. In the third quarter of 2016, Facebook and Twitter had 1.79 billion[1] and 317 million[2] monthly active users, respectively, sharing content about their daily lives and things that happen around them. These systems have revolutionized the way we communicate, by organizing our offline social relationships in a digital form.

Increasingly, our society has been treating actions in the online and offline space in an indistinguishable way. It is now common to see cases in which content posted on Facebook or Twitter is used to vet job candidates, support divorce litigation, or terminate employees Dutta [2010]; Ronson [2016]. Although these platforms provide a democratic space for conversations and debates, they also open space for key unsolved issues of our society, such as gender and race inequalities.

The massive popularity of online social media also provides the opportunity to detect useful characteristics and patterns about users and their interconnections. For instance, patterns are valuable for marketing and advertisement companies which capture users behavior and needs in order to promote products, specifically on a target group. In terms of groups, demographics constitutes a significant factor to cluster people and understand their behavior.

Twitter is a micro-blogging platform in which the main form of communication and interaction is based on posting text messages. It is also a very large system with lots of users. Consequently, it gains the attention of companies and researchers to explore who are these users and what they post on this social network. Although it provides a log of a number of human interaction and social connections, its data is unstructured and lacks information about demographics of users, such as gender, race,

---

[1]https://goo.gl/8GC6Ii
[2]http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

and age. So, the existing efforts that attempt to explore demographics of users on this space rely on different inference strategies. These efforts include Blevins and Mullen [2015]; Karimi et al. [2016]; Liu and Ruths [2013]; Mislove et al. [2011], and Nilizadeh et al. [2016].

Despite the magnitude of Twitter and the amount of the provided information, inferring such demographic aspects of users is challenging and the existing works have several limitations due to the difficulty to identify and gather such demographic aspects. However, the potential to measure demographic aspects is immense and valuable for different research purposes and applications, such as design systems, to understand the social aspects of our society, and how different demographic groups interact with each other. This dissertation focuses on this context. In the next section, we present the main motivations of our work.

1.1   Motivation

Exploring demographic aspects is important both from the perspective of systems and from the sociological point of view. In the first, demographic aspects allow us to understand groups of users that perform certain actions and interactions within a system. As from the sociological point of view, demographic aspects also help us to understand the gender and race differences in online interactions and, consequently, to better understand our society.

Demographic aspects also allow us to study the linguistic differences as well as diversity in topics of interest between gender and race. We believe that studying the differences in how demographic groups write content about specific topics would help systems developers to provide more transparency to users in their systems. The ability of understanding transparency is for algorithms that operate on large-scale crowdsourced data, the goal is to make the outputs of the algorithms (and the algorithm itself) transparent, as it is also important to understand the non-uniformities in the inputs to the algorithms.

Recent efforts show gender disparities on online systems which highlight a concern regarding the inequalities and the main barriers that keep a given demographic from rising beyond a certain level in a hierarchy. For example, women are underrepresented in the content production in Wikipedia. Further, black tenants have much fewer chances of renting a place on Airbnb[3]. Some drivers for both UberX[4] and Lyft[5]

---

[3]http://www.airbnb.com
[4]http://www.uber.com
[5]http://www.lyft.com

discriminate on the basis of the perceived race of the traveler. There is also evidence where disadvantages against female users are found among the users with high visibility Nilizadeh et al. [2016], similar to the barrier females face in attaining higher positions in companies ("glass ceiling" effect) Cotter et al. [2001]. Therefore, identifying inequalities and asymmetries in demographics in the online world is crucial for the development of features that can promote equality and diversity in these systems.

Despite the recent surge of interest in demographic aspects of users from social networks, inferring demographic aspects is challenging and limited due to the difficult to identify and gather such aspects from the users in these systems. The main challenge is that demographic information is not directly available through Twitter API and it needs to be inferred by other means. Hence, in our work, we crawl a large scale sample of active Twitter users and then we identify the gender and race of about 1.6 million users located in the United States by using Face++[6], a face recognition software able to recognize gender and race of identifiable faces in the user's profile pictures. Actually, the state of the art algorithms, for pattern recognition and image processing, can provide, with high accuracy, gender, race, and even the age of an individual via his/her image. In the next section, we discuss our goals.

1.2   Goals

The main goal of this dissertation is the use of demographic aspects along four axes:

- Investigate inequalities in terms of visibility: We are interested in analyzing the association of the demographic aspects with visibility and discovering possible inequalities.

- Investigate linguistic aspects and topic interests: We want to investigate how demographic groups differ from each other in terms of linguistic aspects and topic interests. In other words, we want to find out the extent to which different demographic groups differs from each other in terms the way they post content in Twitter and also the extent to which they have interest in some specific topics.

- Characterize interconnections: We want to examine the proportion of connections and interactions among each demographic group.

- Use demographic aspects to design transparent systems: Demographics is an important aspect of providing transparency. Therefore, we want to design a system to provide demographic transparency of trending topic in Twitter.

---

[6]http://www.faceplusplus.com

In the next section, we present the contributions of this dissertation.

1.3   Contributions

This dissertation presents the following contributions regarding each of our previous goals.

Our first contribution is to provide a clear insight into how groups of users from different demographics connect on Twitter. As an example of finding, we show that white and male users tend to be more followed. Further, our findings reinforce previous observations about disadvantages against female users in Twitter and identify advantages for white users, in comparison with those identified as black and asian. We also show that users perceived as men experience a "glass ceiling" when they are identified as black users.

When we compare the difference between the groups, we find interesting trends, such as the higher interest in zodiac signs by females than by males. Also, white users are more interested in politics, writers, and organizations when compared to asians, and more interested in technology, movie, and politics when compared to black users. On the other hand, black users are more into artists, life, and music topics. We find that asians are more interested in artists, actors, and music than whites and tend to have higher interest for movie, companies, and technology when compared to blacks.

Regarding the interconnection of demographic groups, we found that males and females connections take their part of the responsibility on gender inequality, and that race plays a more important role for connections with white males. Our analysis among these groups explain part of causes of such inequalities and offer hints for the promotion of equality in the online space.

Finally, we believe our findings not only contribute to the social science literature, but we also hope that they can contribute to the recommendation systems design that can act to diminish such inequalities. Thus, we designed the *Who Makes Trends?*[7] Web-based service in order to provide demographic transparency of trending topics in Twitter.

The results presented in this dissertation are part of the following papers (in chronological order of publication):

- **Messias, J.**, Vikatos, P., and Benevenuto, F. (2017). White, Man, and Highly Followed: Gender and Race Inequalities in Twitter. In Proceedings of the

---

[7]http://twitter-app.mpi-sws.org/who-makes-trends

IEEE/WIC/ACM International Conference on Web Intelligence (WI). Messias et al. [2017]

- Vikatos, P., **Messias, J.**, Miranda, M., and Benevenuto, F. (2017). Linguistic Diversities of Demographic Groups in Twitter. In Proceedings of the 28th ACM Conference on Hypertext and Social Media (Hypertext). Vikatos et al. [2017]

- Reis, J. C. S., Kwak, H., An, J., **Messias, J.**, and Benevenuto, F. (2017). Demographics of News Sharing in the U.S. Twittersphere. In Proceedings of the 28th ACM Conference on Hypertext and Social Media (Hypertext). Reis et al. [2017]

- Chakraborty, A., **Messias, J.**, Benevenuto, F., Ghosh, S., Ganguly, N., and Gummadi, K. P. (2017). Who Makes Trends? Understanding Demographic Biases in Crowdsourced Recommendations. In Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM). Chakraborty et al. [2017]

- Kulshrestha, J., Eslami, M., **Messias, J.**, Zafar, M. B., Ghosh, S., Gummadi, K. P., and Karahalios, K. (2017). Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. In Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW). Kulshrestha et al. [2017]

In the next section, we present the dissertation outline.

1.4   Dissertation Outline

This dissertation is organized as follows.

Chapter 2 briefly reviews related efforts in the literature along six axes. First, we discuss the methodology used by efforts that measure demographic aspects in Twitter. Second, we discuss the main efforts that have explored gender and race inequality in social networks. Third, we refer studies that combine linguistic with demographic status. Then, we introduce some efforts regarding algorithmic and data transparency. Moreover, we discuss recommendation diversity. Next, we discuss the main efforts regarding fairness.

In Chapter 3, we present our methodology to gather the demographic data of Twitter users. It consists of gathering tweets of U.S. users and using their profile picture in order to gather their demographic information. We also discuss how we infer the linguistic metrics of tweets using the Linguistic Inquiry and Word Count (LIWC).

Then, we describe our methodology to gather the social connections and interactions between users. Finally, we discuss the potential limitations of our data.

Next, Chapter 4 characterizes gender and race inequalities in visibility by analyzing the association of demographic status with visibility and discovering possible inequalities.

In Chapter 5, we introduce linguistic patterns of users based on gender and race. We also analyze the differences in terms of topics interests.

Then, Chapter 6 discusses results in terms of connections and interactions in Twitter. In other words, we analyze the interconnection among groups of users separated by gender and race and the probability of male and female users to be friends with (i.e. to follow) other male and females user.

Chapter 7 presents a system design of demographic distribution of Twitter trending topics and describes its data collection and trending topic analysis steps.

Finally, Chapter 8 concludes the dissertation and offers directions for future work.

Chapter 2

Related Work

In this section, we review the related literature along six axes. First, we discuss the methodology used by efforts that measure demographic aspects in Twitter. Second, we discuss the main efforts that have explored gender and race inequality in social networks. Third, we refer studies that combine linguistic with demographic status. Then, we introduce some efforts regarding algorithmic and data transparency. Moreover, we discuss recommendation diversity. Next, we discuss the main efforts regarding fairness. Finally, we present the concluding remarks.

2.1   Demographics in Social Media

There are several studies that retrieve information from social media to analyze and predict real-world phenomena such as stock market Bollen et al. [2011], migration Messias et al. [2016], election O'Connor et al. [2010]; Tumasjan et al. [2010], and also political leaning inference of Twitter users Kulshrestha et al. [2017]; Pennacchiotti and Popescu [2011].

One of the first efforts to extract and analyze demographic information presents a comparative study between the demographic distribution of gender/race of Twitter users and U.S. population Mislove et al. [2011]. After that, several efforts have arisen that investigate demographic information, in various social media, using different strategies for distinct purposes Blevins and Mullen [2015]; Karimi et al. [2016]; Burger et al. [2011]; Wachs et al. [2017]; Hannák et al. [2017].

Recent studies focused on demographics Blevins and Mullen [2015]; Karimi et al. [2016]; Liu and Ruths [2013]; Nilizadeh et al. [2016]; Vikatos et al. [2017] present methodologies to extract the necessary data through analysis and pattern matching of screen/full name as well as descriptions of user profiles and image in the profile

status. Particularly, Chen et al. [2015] focus on demographic inference using profile self-descriptions and profile images. They analyze five signals in order to categorize users to their demographic status. They use users' names, self-descriptions, tweets, social networks, and profile images to infer demographic aspects (ethnicity, gender, age). Culotta et al. [2015] describe an alternative methodology assuming that the demographic profiles of visitors to a website are correlated with the demographic profiles of followers of that website on social network (e.g. Twitter), and propose a regression model to determine demographic aspects (gender, age, ethnicity, education, income, and child status). In another study, Reis et al. [2017] study demographic through news sharing in U.S. perspective and show that male and white users tend to be more active in terms of sharing news, biasing news audience to the interest of these demographic groups.

Another study, An and Weber [2016] examine the correlation of hashtags that are used in different demographic groups, using Face++ and users' profile pictures. They show that this strategy is reliable and provides accurate demographic information for gender and race. More recently, Chakraborty et al. [2017] analyze the demographic of Twitter trending topics, using Face++ as well. They use an extensive dataset collected from Twitter in order to make the first attempt to quantify and explore the demographic biases in crowdsourced recommendations. They find that a large fraction of trends is promoted by crowds whose demographics are significantly different from the overall Twitter population. More worryingly, certain demographic groups are systematically under-represented among the promoters of the trending topics.

2.2   Inequality in Twitter Visibility

Homophily states that similarity breeds social connections McPherson et al. [2001]. In our society, people tend to characterize and categorize others in terms of demographic aspects, such as gender and race Willis and Todorov [2006], making them intuitively natural concepts through which people might account similarities. The existing literature is still limited in explaining the inequalities that arise from the establishments of these social connections in Twitter.

Most of the existing efforts in this space attempt to quantify and understand other factors that are highly correlated with visibility (i.e. number of followers). Particularly, Morales et al. [2014] show that user's reputation and her impact on society constitute significant factors of the online visibility growth. Other efforts show that the attention that users gain is correlated with his/her online behavior in terms of the

frequency of their interactions Comarela et al. [2012]; Romero et al. [2011]; Wu et al. [2011]. Freitas et al. [2015] created 120 socialbots in Twitter, with different behaviors and demographics aspects. They show that users who post more, post about specific topics, reply, favorite or mention others frequently, may acquire more followers, and consequently, more visibility. Despite the extreme importance of these efforts, they do not investigate the correlation of demographic issues with visibility. Thus, our effort is complementary to theirs.

There have been efforts quantifying differences and inequalities in many other social media, including Wikipedia Graells-Garrido et al. [2015]; Wagner et al. [2016] and Pinterest Gilbert et al. [2013]. However, the study that is closest related to ours explores gender inequalities in Twitter Nilizadeh et al. [2016]. Authors show that gender may allow inequality to persist in terms of online visibility. This dissertation further explores race as a new demographic dimension. More important, unlike previous studies, our work focuses on investigating how different demographic groups (i.e. male/female, asian/black/white) connect with other.

2.3 Demographics and Linguistic Analysis

In the field of demographics, most studies use linguistic analysis in order to extract useful features for predicting demographic information as gender, race, and age. Burger et al. [2011] produce *ngrams* from users' tweets, description, screen name, and full name, in order to predict Twitter user gender. They conclude that the training of an SVM classifier with the combination of all factors can create an efficient and accurate prediction scheme (92% accuracy) for gender classification. Also, Chen et al. [2015] introduce a similar methodology for predicting gender, ethnicity, and age. However, using ngrams from the social neighbors, including followers and friends, and the distribution of 100 generated topics of LDA algorithm as the input of SVM classifier. In their results, the performance of classification is much lower in terms of ethnicity and age. Gilbert et al. [2013] present an interesting statistical overview in Twitter and Pinterest using textual analysis and comparing what users write on Pinterest to what write text in Twitter. Cunha et al. [2012] used Twitter data to analyze the difference between males and females in terms of generation of hashtags. Their results emphasize gender as a factor that influences the user's choice of specific hashtags to a specific topic.

We motivate this research topic based on De Choudhury et al. [2017] study which discovers gender and cultural differences in Twitter. They correlate several linguistic

features to mental illness. Our findings reinforce their observations about linguistic and topical differences among male and female users in Twitter and also contribute with a new analysis of race Vikatos et al. [2017].

2.4   Algorithmic and Data Transparency

Increasingly, researchers and governments are recognizing the importance of making algorithms transparent. The White House recently released a report that concludes that the practitioners must ensure that AI-enabled systems are open, transparent, and understandable Felton and Lyons [2016]. In news production, black-box automated writing bots or any other difficult-to-parse algorithmic systems used to produce or curate news may represent infringements on the basic ethics of journalism Diakopoulos and Koliska [2016].

Indeed, the controversy around Facebook using human editors on their trending topics teaches a lesson about the importance of transparency. On one hand, when humans were editing trending topics, they were accused to select and filter content Nunez [2016]. On the other hand, when humans were removed from the process, Facebook was accused of featuring fake news as trending Ohlheiser [2016]. Recently, Twitter announced the use of deep learning in their timeline algorithm and how it works in order to make it transparent to the public Koumchatzky and Andryeyev [2017].

Demographics is also an important aspect of providing transparency. Therefore, demographic distribution of users may be leveraged to make other crowdsourced systems (e.g., social search Kulshrestha et al. [2017] and Twitter trending topic analysis Chakraborty et al. [2017]) transparent as well. For algorithms that operate on large-scale crowdsourced data, it is paramount to make the outputs of the algorithms (and the algorithm itself) transparent, and understand the non-uniformities in the inputs to the algorithms.

2.5   Recommendation Diversity

Diversity is increasingly recognized as an important metric for evaluating the effectiveness of online recommendations Chen et al. [2016]. There have been multiple attempts in the past to introduce diversity in content search and recommendation systems. Gollapudi and Sharma [2009] and Agrawal et al. [2009] explored several approaches for search result diversification. Ziegler et al. [2005] proposed algorithms to introduce *Novelty and Serendipity* in collaborative filtering based recommendations. Basically,

search result diversification suggests a trade-off between showing diverse results in the top position for a given query and have more appropriate results that approximate to the user intent Carbonell and Goldstein [1998]; Chen and Karger [2006]. There is also an effort to study stereotypes in Google and Bing search engines. Araújo et al. [2016] query the search engines for beauty and ugly females. Their results show the existence of both negative stereotype for black females and positive stereotype for white females in terms of beauty. They also noticed that there are negative stereotypes about older females in U.S. while in hispanic countries share negative stereotypes about black females, positive for whites and neutral for asians. Sheth et al. [2011] propose the "Social Diversity", a new approach to diversity in recommendations that uses social network information to diversify recommendation results.

In 2015, Google search engine suffered from a lack of diversity when their search algorithm labeled black people as "gorillas" Guynn [2015]; Barr [2015]. It reinforces the importance of using demographic information to provide diversity in recommendation systems. More recently, in 2016, a Microsoft Twitter bot that uses artificial intelligence quickly became a racist Victor [2016]. It also highlights the need for a better understanding of designing systems. Depending on which kind of output the algorithm gives to someone it could ruin her life.

2.6 Fairness

Some countries have anti-discrimination laws in order to prohibit unfair treatment of individuals based on sensitive attributes (e.g. race and gender). It can be hard to identify the source of the problem or to explain it to a court Barocas and Selbst [2016]. Moreover, in 2013 the New York Police Department (NYPD) started the Stop-question-and-frisk program (SQF[1]) to reduce criminality. It consists of temporally detaining, questioning, and at times searching civilians on the street for weapons and other contraband. Goel et al. [2016] found that blacks and hispanics make up more than 80% of individuals stopped by SQF, even though they constitute approximately 50% of the New York City population.

Likewise, labor discrimination is also a difficult problem that has been studied for many years Todisco [2014]. There are some studies that uncovered discrimination in traditional labour markets Bertrand and Mullainathan [2004]; Carlsson and Rooth [2007]; Dovidio and Gaertner [2000]. More recently, Hannák et al. [2017] explore the

---

[1]https://en.wikipedia.org/wiki/Stop-and-frisk_in_New_York_City

correlations between two freelance marketplaces, TaskRabbit[2] and Fiverr[3], and quantify race- and gender-biases. They found that on TaskRabbit females, especially white females, receive 10% fewer than males with equivalent work experience. Black workers, especially males, receive significantly lower feedback scores than others. On the other hand, on Fiverr, black workers, especially males, receive 32% fewer reviews than other males. Asian workers, also especially males, receive higher scores than others. Finally, their linguistic analysis shows that reviews for black females include fewer positive adjectives, while for blacks (in general) more negatives. Another, study shows that black tenants have much fewer chances of renting a place on Airbnb[4] Edelman et al. [2017]. Also, some drivers for both UberX[5] and Lyft[6] discriminate on the basis of the perceived race of the traveler Ge et al. [2016]. Google served ads that despised African-Americans Sweeney [2013], whilst Google did not show ads for high-paying jobs from females Datta et al. [2015]. Further, black and female sellers earn less than white and male sellers on eBay[7] Ayres et al. [2015]; Kricheli-Katz and Regev [2016].

A number of studies found racial disparities in automated Angwin et al. [2016] as well as human Goel et al. [2016]; Ridgeway and MacDonald [2009] decision making systems related to criminal justice. Recent studies focus on designing decision making systems that avoid unfairness Feldman et al. [2015]; Luong et al. [2011]; Zafar et al. [2017b]; Zemel et al. [2013], proposing methods for detecting Feldman et al. [2015]; Luong et al. [2011]; Pedreshi et al. [2008]; Romei and Ruggieri [2014] and removing Dwork et al. [2012]; Feldman et al. [2015]; Goh et al. [2016]; Kamiran and Calders [2012]; Kamishima et al. [2012]; Pedreshi et al. [2008]; Zafar et al. [2017b]; Zemel et al. [2013] unfairness. Furthermore, Zafar et al. [2017a] propose fair classifier formulations in order to remove disparate only on false-positive and false-negative rates. They provide a flexible trade-off between disparate mistreatment-based fairness and accuracy.

Demographics is key to providing more fairness algorithm and classifiers. Therefore, we hope our study may help developers and the research community to improve their systems in order to provide tools to decrease racial and gender disparities.

---

[2] https://www.taskrabbit.com
[3] https://www.fiverr.com
[4] http://www.airbnb.com
[5] http://www.uber.com
[6] http://www.lyft.com
[7] http://www.ebay.com

2.7 Concluding Remarks

In this chapter, we provided related work regarding the importance of demographic aspects of users in Online Social Networks (OSN).

Our effort uses a similar strategy to gather demographic information as Chakraborty et al. [2017], but we investigate very different research questions as we focus on inequalities, demographic aspects of users connections, and linguistic patterns of users. We note that our strategy to gather demographic aspects, applied to active users within the United States, allows us to take a step further in many existing studies on the field. It allows us to study race inequalities and also allows us to investigate how connections are established among demographic groups (i.e. male/female, asian/black-/white). We also quantify to what extent one group follows and interacts with each other and the extent to which these connections and interactions reflect in inequalities in Twitter.

In terms of data transparency, we hope that the methodology to compute demographic distribution of users can be leveraged to make other crowdsourced systems (e.g., social search Kulshrestha et al. [2017] and Twitter trending topic analysis Chakraborty et al. [2017]) transparent as well. More importantly, for algorithms that operate on large-scale crowdsourced data, is to make the outputs of the algorithms (and the algorithm itself) transparent, as it is also important to understand the non-uniformities in the inputs to the algorithms.

Regarding fairness, we hope our study may improve systems in order to provide tools to decrease racial and gender disparities. Demographic information is also an important factor to take in consideration to improve diversity in content search and recommendation system. Researchers could use age, gender, and race as attributes to provide diverse results to users.

Demographics also reflect users' intent which may improve search engines results. All of the existing techniques of diversification have focused on bringing in diversity in the *topical coverage of the contents*. Here, we propose a fundamentally new approach to introducing diversity. Instead of focusing on the diversity of content coverage, we attempt to increase diversity in the population who promotes the contents Chakraborty et al. [2017]. Therefore we developed a system to show the demographic distribution of users in Twitter. We present this system in Chapter 7.

We hope our new and large-scale dataset may largely contribute to the research community. In the next chapter, we introduce the data collection of this dissertation.

Chapter 3

Demographic Information Dataset

This chapter describes the data collection and methods to infer demographic information of individual Twitter users. Our ultimate goal consists of gathering a dataset containing active U.S. Twitter users, their demographic information (gender and race), a sample of their connection graph (friends), and their tweets. Next, we describe our steps to create two distinct datasets, (i) inequalities in visibility and linguistic patterns of users; (ii) social connections and interactions of users; and also discuss their main limitations.

3.1 Twitter Dataset Gathering

To identify active Twitter users we gathered data from the 1% random sample of all tweets, provided by the Twitter Stream API [1], it offer samples of the public data flowing through Twitter. Our dataset covers a period of three complete months, from July to September 2016. In total, we collected $341,457,982$ tweets posted by around $50,270,310$ users.

We restricted our demographic studies to Twitter users located in the United States who posted during this period, at least, one tweet in English. As geographic coordinates are available in Twitter only for a limited number of users (i.e. $< 2\%$) Cha et al. [2010], our strategy to identify U.S. Twitter users was based on part of the methodology used in previous efforts Kulshrestha et al. [2012]; Chakraborty et al. [2017]. We have used the time zone information as a first filter and then we attempted to remove from this filtered dataset those users that provided free text location indicating they are not U.S. (i.e. Montreal, Vancouver, Canada). We end up with a dataset containing $6,286,477$ users likely located in the United States.

---

[1] https://dev.twitter.com/streaming/public

3.2   Crawling Demographic Information

Most of the existing studies related to demographics of users in Twitter have looked into gender and age. Some efforts attempt to infer the user's gender from the user name Blevins and Mullen [2015]; Karimi et al. [2016]; Liu and Ruths [2013]; Mislove et al. [2011]. However, some users may not use their real names, consequently, their gender could not be inferred properly Liu and Ruths [2013]. Others have attempted to identify patterns like '25 *yr old*' or '*born in* 1990' in Twitter profile description to identify the user age Sloan et al. [2015].

From here, we use a different strategy that allows us to study another demographic aspect: the user's race. To do that, we crawl the profile picture Web link of all Twitter users identified as located within the United States. In December 2016, we crawled the profile picture's URLs of about 6 million users, discarding $4,317,834$ (68.68%) of them. We discarded users in two situations: first, when the user does not have a profile picture and second, when the user has changed her picture since our first crawl. When users change their picture, their profile picture URL changes as well, making it impossible for us to gather these users in a second crawl via Twitter Stream API.

From the remaining $1,968,643$ users, we submitted the profile picture Web links into the *Face++ API*. Face++ is a face recognition platform based on deep learning Fan et al. [2014]; Yin et al. [2015] able to identify the gender (i.e. male and female), age, race (limited to asian, black, and white), and more attributes related to smiling, face positions, glasses information from recognized faces in images. In this dissertation, we focus on gender and race attributes.

We have also discarded those users whose profile pictures do not have a recognizable face or have more than one recognizable face, according to Face++. Our baseline dataset, also used by Chakraborty et al. [2017], contains $1,670,863$ users located in U.S. with identified demographic information. Table 3.1 shows the demographic distribution of users in our baseline dataset across the different demographic groups. The phases of our data crawling and the amount of data on each step are summarized in Table 3.2.

Table 3.1: Absolute demographic distribution of 1.6 million users in baseline dataset.

Race	Gender		Total
	Male	**Female**	
Asian	$120,950$ (7.24%)	$177,205$ (10.61%)	$298,155$ (17.85%)
Black	$130,954$ (7.84%)	$107,827$ (6.45%)	$238,781$ (14.29%)
White	$538,625$ (32.23%)	$595,302$ (35.63%)	$1,133,927$ (67.86%)
Total	$790,529$ (47.31%)	$880,334$ (52.69%)	$1,670,863$ (100%)

Table 3.2: Dataset construction

Phase	Number of Users
Crawling 3 months of Tweets	50 million
Filtering U.S. users	6 million
U.S. users with profile image	1.9 million
Baseline: U.S. users with one face	1.6 million
Dataset 1: Recognized U.S. users with linguistic attributes	$304,477$
Dataset 2: Social connections and interactions	$428,697$

3.3 Baseline Dataset

In this section, we use a null model as a method to estimate the statistical significance of the observed trend in the given data. We compare the distribution of random samples created by the null model with that of the original dataset and we measure the statistical significance. This step is important due to some limitations to gather all tweets and linguistic attributes for 1.6 million users. Therefore, to circumvent this limitation, we prefer to randomly select $304,477$ users from the 1.6 million dataset and show the statistical significance of this sub-dataset which we use to show results regarding race and gender inequalities and linguistic patterns of Twitter users.

Table 3.1 shows the distribution of gender and race in the baseline dataset of the $\approx$ 1.6 million Twitter users between July and September 2016. Additionally, Table 3.3 shows the demographic breakdown of the $304,477$ randomly selected users across the different demographic groups. To construct a null model, we create k random samples from the whole population (our crawled dataset with demographic aspects), where each sample has exactly $304,477$ users. For each sample, we count how many asians, blacks, whites, males, and females are included. Then, the Z_{White} is computed as follows:

$$Z_{White} = \frac{|U_{White}| - mean(|S_{White}|)}{std(|S_{White}|)} \qquad (3.1)$$

where $mean(White)$ is the mean and $std(White)$ is the standard deviation of the values from multiple samples of white users. We use the same equation for the other gender and race aspects. Table 3.4 presents a detailed description of the comparison with null model. Additionally, Table 3.3 presents the demographic distribution of $304,477$ users with linguistic attributes. The numbers in the parenthesis correspond to the Z-values.

Intuitively, when the absolute value of Z-value becomes bigger (either positive or negative), the trend (more or less number, respectively) is less likely observed by

Table 3.3: Demographic distribution of $304,477$ users with linguistic attributes. The numbers in the parenthesis correspond to the Z-values.

Race (%)	Gender (%)		Total (%)
	Male	**Female**	
Asian	7.07 (-3.85)	10.05 (-11.28)	17.12 (-10.90)
Black	8.17 (8.53)	6.74 (7.68)	14.91 (11.69)
White	32.88 (8.49)	35.09 (-7.69)	67.97 (1.20)
Total	48.12 (10.91)	51.88 (-10.91)	100.00

Table 3.4: Basic statistical description of null models. $k = 100$ samples with a population of $304,477$ randomly selected users. We use confidence intervals of 95% confidence level.

Demographic	Mean	Z-value	S.D.	Min	25-perc	Median	75-perc	Max
Male	$144,035.1 \pm 44.86$	10.91	228.88	143,544	143,883.00	144,054.5	144,156.50	144,680
Female	$160,441.9 \pm 44.86$	-10.91	228.88	159,797	160,320.50	160,422.5	160,594.00	160,933
Asian	$54,311.5 \pm 39.17$	-10.90	199.87	53,907	54,177.25	54,296.5	54,444.00	54,803
Black	$43,514.01 \pm 31.72$	11.69	161.85	43,196	43,380.75	43,503.5	43,633.50	43,887
White	$206,651.49 \pm 46.82$	1.20	238.91	205,921	206,490.25	206,666.5	206,789.25	207,110
Asian Male	$22,043.64 \pm 26.24$	-3.85	133.88	21,674	21,958.75	22,040.5	22,115.50	22,429
Asian Female	$32,267.86 \pm 28.92$	-11.28	147.56	31,900	32,153.50	32,262.0	32,371.75	32,667
Black Male	$23,857.98 \pm 23.81$	8.53	121.48	23,634	23,777.75	23,858.0	23,930.00	24,197
Black Female	$19,656.03 \pm 21.82$	7.68	111.34	19,342	19,585.25	19,660.5	19,737.75	19,944
White Male	$98,133.48 \pm 45.61$	8.49	232.73	97,538	97,995.25	98,130.5	98,297.50	98,623
White Female	$108,518.01 \pm 43.04$	-7.69	219.62	108,025	108,348.25	108,501.5	108,688.00	109,015

chance. In this work, the size of population is $304,477$, and $k=100$.

3.4   Gathering Tweets

We are interested in correlating linguistic features of Twitter users with demographic information. We crawled the recent $3,200$ tweets of $304,477$ users for the purpose of linguistic analysis. Table 3.3 shows the demographic breakdown of users in this dataset across the different demographic groups. We can note a prevalence of females (51.88%) in comparison to males (48.12%) and a predominance of whites (67.97%) in comparison to blacks (14.91%) and asians (17.12%). This means if we pick users randomly in our dataset, we would expect demographic groups with these proportions. Table 3.5 shows the statistical descriptions of the number of tweets with confidence intervals of 95% confidence level for each demographic groups.

Table 3.5: Basic statistical descriptions of number of tweets with confidence intervals of 95% confidence level.

Demographic	Mean	Median	Max
Male	$11,624.76 \pm 109.40$	$3,874$	$1,683,948$
Female	$12,933.40 \pm 105.89$	$4,885$	$1,132,964$
Asian	$14,020.92 \pm 183.73$	$5,544$	$1,108,525$
Black	$18,949.91 \pm 248.46$	$8,245$	$973,225$
White	$10,432.49 \pm 85.28$	$3,637$	$1,683,948$

3.5 Extraction of Topics

We extracted the information about topics of interests for active users using the Who Likes What[2] Web-based service Bhattacharya et al. [2014]. The produced topics are derived from the list of friends (other users the user is following) of each user. Then, we sort the produced topics based on their frequency to obtain the 20 most common topics from Twitter users, including them as binary variables (1 if the user is interested in this topic, 0 otherwise). We manually cleaned several top topic labels following the same procedure as Nilizadeh et al. [2016]. Therefore, we merge topics like *businesses* and *biz*, group topics into similarity (e.g. *celebrities* and *famous*, *actors* and *actor*), and remove some topics like *best*, *br*, *bro*, *new*. Table 3.6 presents a list of the 20-top topics and the merged sub-topics in each one as well as the number of users that are interested in.

3.6 Linguistic Measures

To quantify gender and race dimensions in the language of Twitter users, we use the psycholinguistic lexicon Linguistic Inquiry and Word Count (LIWC) Tausczik and Pennebaker [2010]. The crawled tweets of a user were gathered in a supertext (the concatenation of all tweets) in order to extract a variety of linguistic metrics. The features are categorized into 3 main categories, (1) affective attributes; (2) cognitive attributes; and (3) linguistic style attributes, as De Choudhury et al. [2017] propose. For this work, we considered 36 features from LIWC categorized in six groups (affective attributes, cognitive attributes, lexical density and awareness, temporal references, social/personal concerns, and interpersonal focus) in order to find the main differences across each demographic group.

---

[2]http://twitter-app.mpi-sws.org/who-likes-what

Table 3.6: 20-top topics of user's interests

Topic	Sub-Topics	#Users
World	world, earth, hollywood, usa, canada, texas, international, nyc, country, city, boston, san francisco, france, america, los angeles, brasil, london, india	290,030
Celebrities	celebrities, famous, stars, celebs, celebrity, star, celeb	245,125
Entertainment	entertainment	244,956
Music	music, pop, hip hop, rap, gospel, hiphop	227,986
TV	tv, television	225,682
Life	life, lifestyle, health, healthcare, fitness, food, style, smile, drink	157,032
Fun	fun, funny, humor, lol, laugh	154,058
Info	info, information	147,567
Artists	musicians, singers, artist, singer, musician, rappers, bands	141,519
Actors	actors, actresses, actress, actor	140,647
Media	sports news, tech news, newspapers, music news, breaking news, world news, news media, radio, internet, social media, youtube, sports media, magazines, magazine	135,849
Writers	writers	126,051
Bloggers	bloggers, blogs, blog	110,699
Business	business, biz, businesses	107,361
Sports	sports, football, basketball, baseball, soccer, futbol, basket, martial arts, sport, mma, golf, cricket, boxing, motorsports, f1, racing	93,611
Movie	movie, movies, film, films	88,863
Organizations	organizations, nfl, nba, mlb, nhl, ufc, lfc, lgbt	82,568
Technology	technology, tech, iphone, digital, geek, software, computer, electronic, android, xbox, mac, gadgets, programming, geeks	72,137
Politics	politics, government, political, politicians, politician	64,735
Companies	companies, apple, company, microsoft, google	53,128

The affective attributes contemplate features that show how strong is the expression of feelings like anger, anxiety, sadness, and swear. Cognitive attributes are related to the process of knowledge acquisition through perception. The lexical density and awareness group gather features related to the language itself and its structure. Temporal references are related to the tense expressed in the writing, while interpersonal focus presents features related to the speech. The social/personal concerns group comprises features that express characteristics inherent to the individual as well his/her relation to the environment where he/she lives.

3.7 Gathering Social Connections and Interactions

Ideally, to study how different demographic groups are connected and interact with each other, we would like to have at our disposal the followers and friends of all users from our dataset. However, the number of followers and friends of the 1.6 million users in our dataset surpasses 6.4 billion users, making it unfeasible to crawl all profiles through the Twitter and the Face++ API. Next, we discuss a sampling procedure we used to create a dataset of users to study their connections and their interactions with demographic information.

We randomly select a total of $6,000$ users from our dataset, $1,000$ users of each demographic group (i.e. asian male, asian female, black male, black female, white male, and white female). Then, we gathered their friends. As some of these users have a prohibitively large number of friends, we limited our crawl to gather only the most recent $5,000$ friends of a given user, as this is the maximum number of user IDs the Twitter API returns per request. However, this strategy gathered the entire friends' list for 98.51% of the users.

Then, we follow the same methodology we discussed before to gather the demographic information of users. First, we remove users that are not located in U.S., and then we attempt to identify the demographic aspects of each user using Face++ API. It is undeniable that the aggregated number of friends is extremely large and it is difficult to gather due to the limitation on the number of requests that Face++ API allows us to make. Thus we were limited to gather demographics of at least 5% of each user's friends.

Our social connections and interactions dataset contains $428,697$ users with the proper demographic information identified. Table 3.7 presents the total number of friends gathered for each demographic group that we examined. The average and median percentage of friends of the $6,000$ users for which we were able to gather

Table 3.7: Number of Friends in each Group

	White	Black	Asian	Total
Male	$151,840$	$52,437$	$24,299$	$228,576$
Female	$137,010$	$31,011$	$32,100$	$200,121$
Total	$288,850$	$83,448$	$56,399$	$428,697$

Table 3.8: Number of Interactions in each Group

	White	Black	Asian	Total
Male	$246,879$	$109,744$	$51,370$	$407,993$
Female	$202,338$	$60,108$	$71,137$	$333,583$
Total	$449,217$	$169,852$	$122,507$	$741,576$

demographic information are 10.15% and 9.40%, respectively. We note that these fractions are usually higher than 5% as some extra users were previously gathered in our 1.6 million demographic dataset.

Also, we worked in a similar way for the analysis of interactions between demographic groups. More specifically, we select the users that retweet and mention the tweets of the randomly selected users. To do that, we gathered all tweets (limited to a maximum of $3,200$ tweets due to Twitter REST API limitation[3]), it returns a collection of the most recent tweets posted by the user, of our $6,000$ users. We then identified which of these users were mentioned or retweeted and we limited to gather the demographic information of only 5% of retweeters and mentioned users for our analysis. Table 3.8 summarizes the amount of crawled users for each demographic group.

Our study about the connections among demographic groups (Chapter 6) is based on this specific dataset.

3.8   Potential Limitations

The gender and race inference are challenge tasks, and as other existing strategies have limitations and the accuracy of Face++ inferences is an obvious concern in our effort. The limitations of our data is discussed next.

Accuracy of the inference by Face++: Face++ itself returns the confidence levels for the inferred gender and race attributes, and it returns an error range for inferred age. In our data, the average confidence level reported by Face++ is $95.22 \pm 0.015\%$ for gender and $85.97 \pm 0.024\%$ for race, with a confidence interval of 95%. Recent efforts have used Face++ for similar tasks and reported high confidence

---

[3]`https://dev.twitter.com/rest/reference/get/statuses/user_timeline`

in manual inspections An and Weber [2016]; Zagheni et al. [2014]; Chakraborty et al. [2017]. In addition, in recent scientific effort Chakraborty et al. [2017] they evaluated the effectiveness of the inference made by Face++, using human annotators to label randomly selected profile images from Twitter. They measured the inter-annotator agreement in terms of the Fleiss' κ score which was 1.0 and 0.865 for gender and race, respectively.

Our dataset may also include fake accounts and bots. Previous studies provide evidence for an significant fraction of fake accounts Freitas et al. [2015]; Messias et al. [2013] in Twitter.

Data: Finally, we note that our approach to identify users located in U.S. may bring together some users located in the same U.S. time zone, but from different countries. We, however, believe that these users might represent a small fraction of the users, given the predominance of active U.S. users in Twitter Cheng et al. [2009]. Also, we are using the 1% random sample of all tweets. However, the 1% random sample is not the best data to capture all the dynamics happening in Twitter, its limitations are known Morstatter et al. [2014] and it is the best available option at our disposal.

3.9 Concluding Remarks

In this chapter, we presented our methodology to gather the demographic data of Twitter users. First, we crawled the English tweets from U.S. in order to get the users who posted them. Second, we submitted their profile picture Web links into the Face++ API. Face++ is a face recognition platform based on deep learning able to identify the gender, race, and age attributes. We also discussed how we inferred the linguistic metrics of tweets using the LIWC. Then, we described our methodology to gather the social connections and interactions between users. Finally, we discussed our potential limitations. In the next chapter, we analyze the association of demographic status with visibility and discovering possible inequalities in Twitter. In other words, we analyze whether the gender and race affect users' number of followers and how many lists they are added to.

Chapter 4

Inequality in Visibility

We are interested in analyzing the relationship between demographic aspects and visibility, and discovering possible inequalities. These asymmetries can be derived from the prejudices and stereotypes in the selection of which user to follow based on gender or race. Nilizadeh et al. [2016] provide evidence of inequalities and asymmetries in terms of visibility between males and females.

We focus only on the visibility in the social network and not in user's influence of the audience to which he/she is exposed. For our purpose, we use two different features to measure a user visibility: its follower count and its listed count. In other words, we test whether the gender and race affect users' number of followers and how many lists they are added to.

The number of followers measures the real size of the audience that someone is exposed Cha et al. [2010]. Finally, listed count represents the number of times a user was added to a specific list by others in Twitter. Twitter Lists allow users to group and organize Twitter accounts that tweet on a topic that is of interest to her and follow their collective tweets. Many users carefully create lists to include other Twitter users who they consider experts on a given topic. Previous research efforts Sharma et al. [2012] attempt to gather lists in a large scale to find experts in Twitter.

As a baseline for comparisons, Table 3.3 shows the demographic breakdown of users in our dataset across the different demographic groups. We can note a prevalence of females (51.88%) in comparison to males (48.12%) and a predominance of whites (67.97%) in comparison to blacks (14.91%) and asians (17.12%). This means that if we pick users randomly in our dataset, we would expect demographic groups with these proportions. We used these proportions next as a baseline for characterizing inequalities in Twitter.

4.1   Gender Inequality

We begin our analysis by sorting all users with respect to the number of followers in
our dataset and then we computing the fraction of males and females among the top
followed users. Figure 4.1 provides some insight on this point. Although, the number
of females (51.88%) is larger than the males (48.12%) as shown in Table 3.3, males
tend to be among the top followed users. The most significant difference exists in top
1% with a fraction of 57.34% of males and 42.65% of females, which is almost 15% and
this difference is decreasing until the top 14%, where the fraction of females become
higher than the fraction of males.

The same kind of observation can be made for the ranking of listed counts, as
shown in Figure 4.2. However, the discrepancy perceived is smaller in the lists. Basi-
cally, it is restricted to the top 1% of the most listed users with 51.23% of males and
48.64% of females. In the top 6%, the fraction of females become higher than males.
The same occurs in the top 8%.
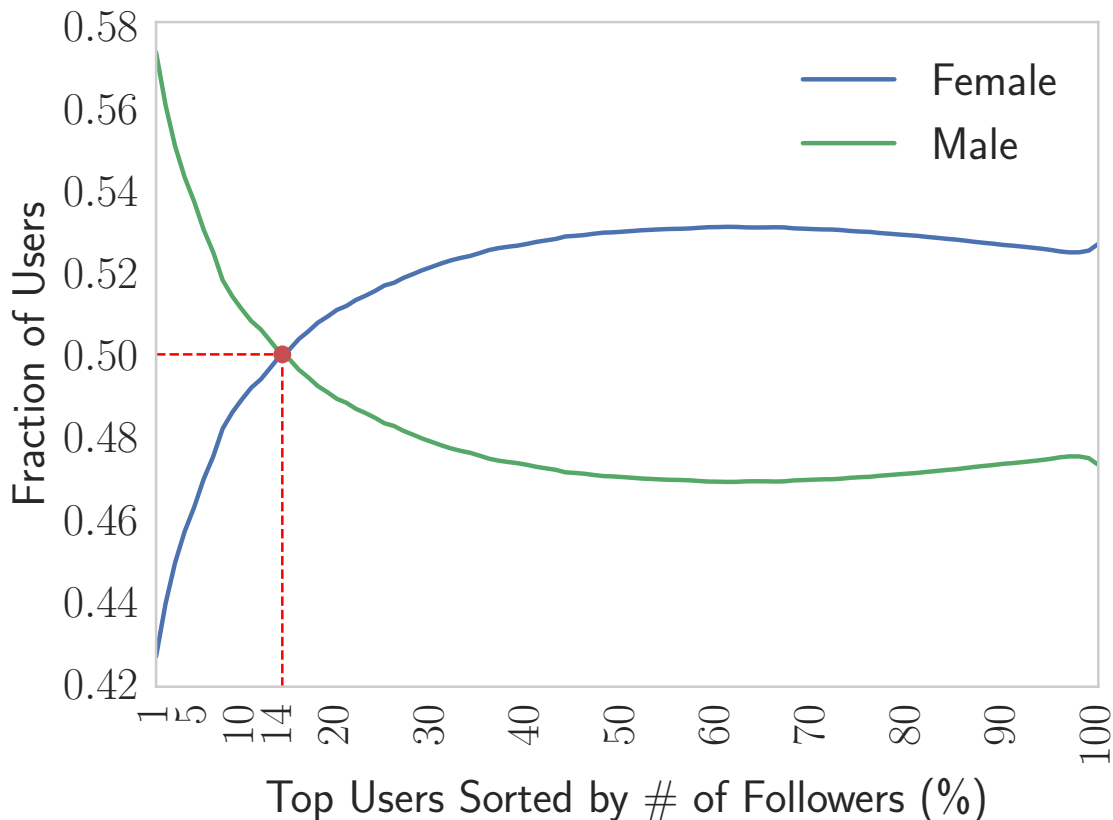
Figure 4.1: Distribution of the fraction of females (in blue color) and males (in green
color) in the rank of users with most followers. The intersection point happens in the
top 14%, meaning the fraction of females became higher than males from this point.
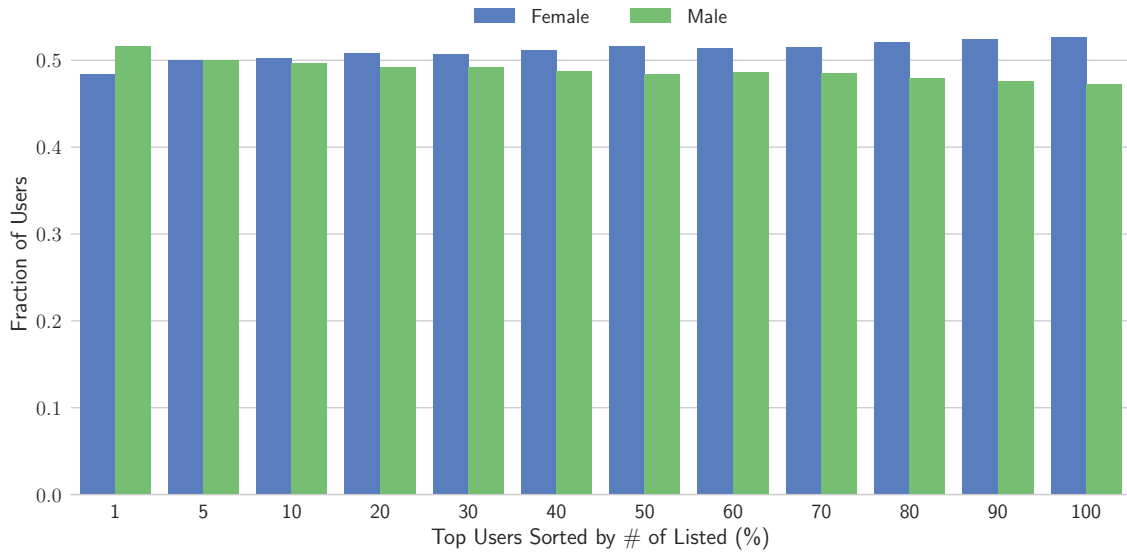
Figure 4.2: Distribution of the fraction of females (in blue color) and males (in green color) in the rank of most listed users.

The results of this analysis reflect the general idea that higher positions are usually taken by males Sugimoto et al. [2013].

Our observations also reinforce previous findings related to gender in Twitter Nilizadeh et al. [2016], suggesting the existing of the so called 'glass ceiling effect', which posits that females face an invisible barrier at the highest levels of an organization Cotter et al. [2001]. One concern raised by authors on that work is that disparity may be driven by a small number of 'elite' users. Our findings diminish such concerns as our dataset is at least two orders of magnitudes larger and we noted such discrepancies even in the top 10% positions.

4.2   Race Inequality

We now turn our attention to the analysis of race inequalities. Similarly to what we have done in the previous section, we examine here the presence of each user race within the top positions in a rank of top followed users and top listed users. Figure 4.3 and Figure 4.4 provide the necessary information of the top followed and the top listed users for each race. We note that the number of white users tends to be higher in the both top positions (in the rank of users with more followers and in the rank of most listed users).

Our results suggest that race disparity in Twitter visibility occurs, meaning that at the highest levels of visibility, users perceived to be white come out on top position.
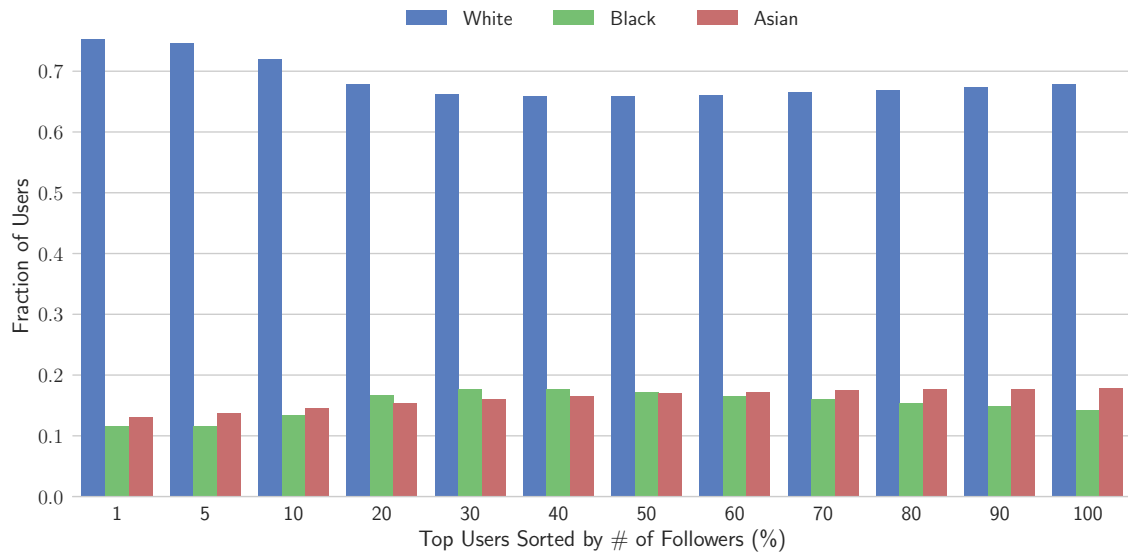
Figure 4.3: Amount of whites, blacks, and asians in the rank of users with most followers
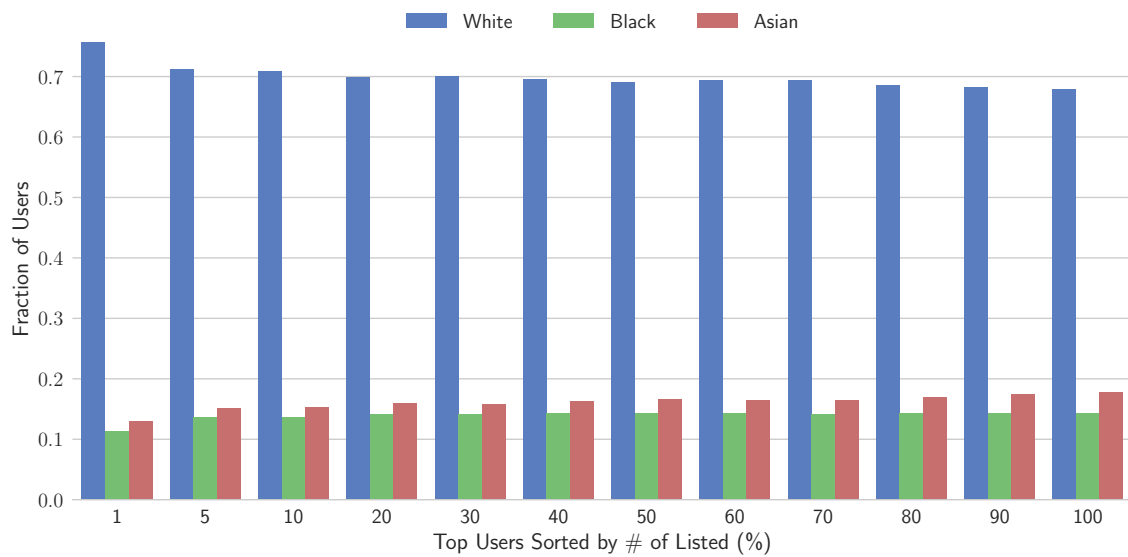
Figure 4.4: Amount of whites, blacks, and asians in the rank of the most listed users

This observation reinforces many previous observations related to race inequality in United States Bonilla-Silva [2006]; Oliver and Shapiro [2006].

4.3   Taking Together Gender and Race Inequality

Finally, we attempt to quantify the existing inequality when we look into gender and race aspects at the same time. We note from Figure 4.5 that the amount of white males

among the top 1% users with more followers is 42.27%. This represents an increase of 9.39% in the fraction of white males in comparison to the expected proportion of white males in our dataset, which is 32.88%. We can note that the population of black females and asian females experience a reduction of 3.9% and 3.29% in the top 1% most followed users, from 6.74% and 10.05% to 2.84% and 6.76%, respectively. We note similar trends in terms of lists as we can see Figure 4.6.
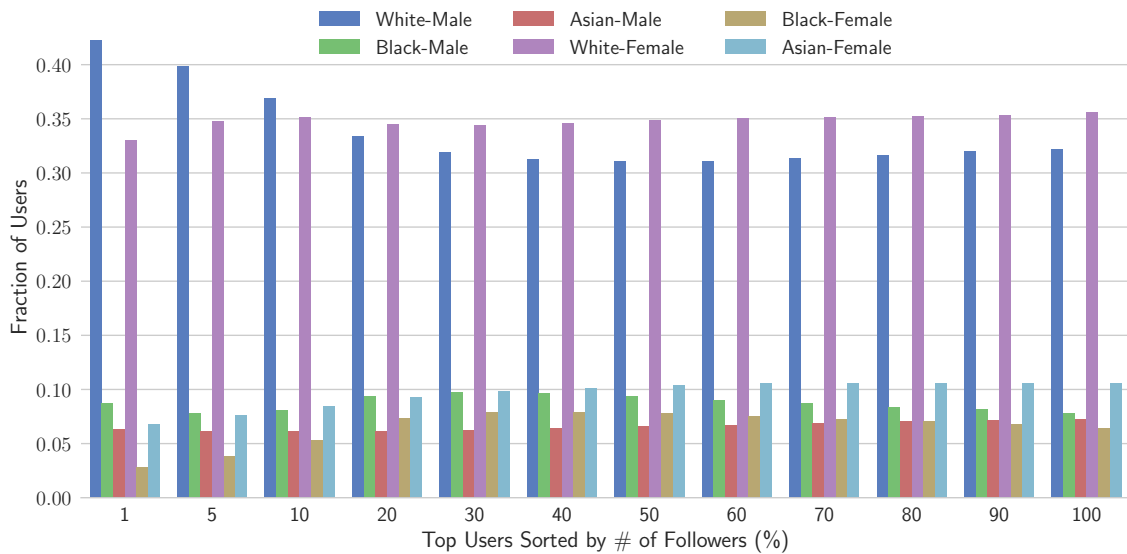
Figure 4.5: Demographic groups in terms of gender and race as a function of the rank of the most listed users

To better analyze to what extent demographic groups in the top 1% is substantially higher or lower than what one would expect based on our demographic information dataset, Table 4.1 shows the increase or decrease in the proportion of each demographic group in comparison with such baseline expected population. Overall, these results quantify the perceived inequalities that six demographic groups experience at the highest levels of visibility (i.e. top 1% most followed and listed). The most privileged demographic group in the rank of followers is the group of white males, with 28.56% higher than we expected for the most followed users. The number of white females is also higher than we expected for the most listed users, but much smaller amount (only 5.04%) in comparison to males. The most unprivileged groups are asian females and black females, under-represented in the top 1% of the most followed users by more than 32%.
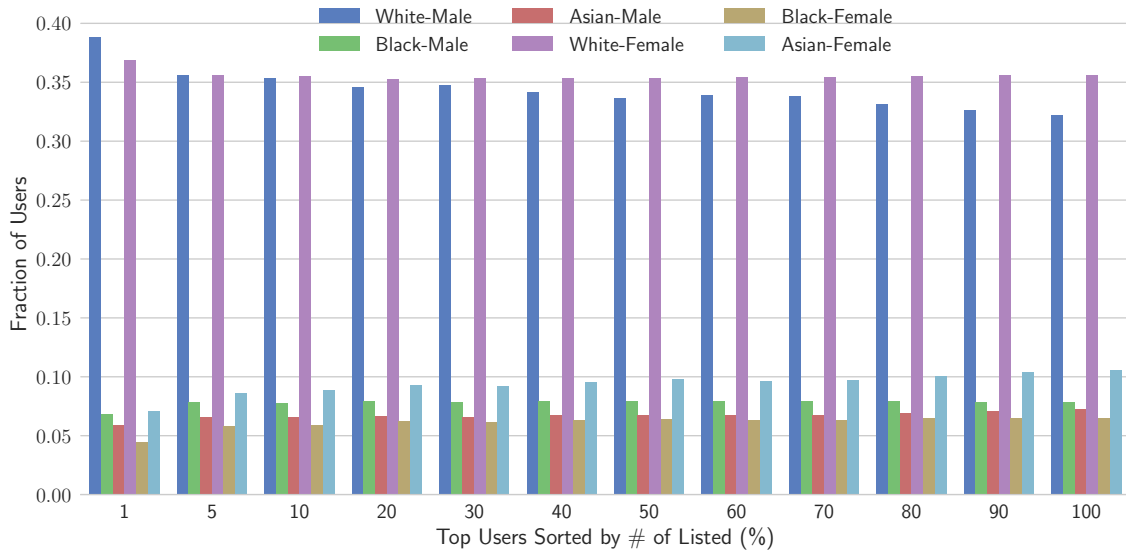
Figure 4.6: Demographic groups in terms of gender and race as a function of the rank of the most listed users

Table 4.1: Relative proportion of each demographic group in the top 1% rank of users with more followers (followers column) and the most listed (listed column) in comparison to the baseline expected population.

Race	Followers		Listed	
	Male	Female	Male	Female
Asian	−10.60	−32.70	−16.36	−29.61
Black	+7.17	−57.73	−15.90	−34.20
White	+28.56	−5.84	+18.15	+5.04

4.4   Concluding Remarks

In this chapter, we analyzed the association of demographic status with visibility and discovering possible inequalities. We showed that the Twitter glass ceiling effect, typically applied to females, also occurs in Twitter for males, if they are black or asian. Our analysis reinforces evidence about gender inequality in terms of visibility and introduces race as a key demographic aspect, which reveals hiding prejudices between demographic groups. In the next chapter, we investigate how demographic groups differ from each other in terms of linguistic and topic interests.

Chapter 5

Linguistic Patterns

In this chapter, we want to investigate how demographic groups differ from each other in terms of linguistic and topic interests. In other words, we want to find out the extent to which male differs from female in terms of the way they post content in Twitter and also the extent to which they have interest for some specific topics. This chapter also presents information regarding race, and combinations of gender and race.

We believe that studying the differences in how demographic groups write content about specific topics would help systems developers to provide more transparency to users. Therefore, first, we present results regarding the linguistic differences of demographic groups. After words, we show their interests in Twitter.

5.1 Linguistic Differences

In order to show how demographic groups differ from each other in both gender and race domains, this section presents the absolute difference between groups with respect to various categories of linguistic measures. Table 5.1 shows the linguistic features extracted from LIWC into six categories (affective attributes, cognitive attributes, lexical density and awareness, temporal references, social and personal concerns, and interpersonal focus).

Figure 5.1 shows the mean absolute differences between male and female users across each linguistic category. The difference for a specific group of features is calculated by taking the average ratio of the difference between the values for male and female to the values of the measure among male. The mean difference in the first group (affective attributes), for example, is calculated as the average of the absolute difference of each feature that comprises this group. It shows in which linguistic category

the analyzed users differ the most. The number of users considered in each group were the same.

Figure 5.1 also shows that interpersonal focus, which contemplates features like family, friends, health, religion, body, achievement, home, and sexual as the most prominent linguistic difference among males and females.
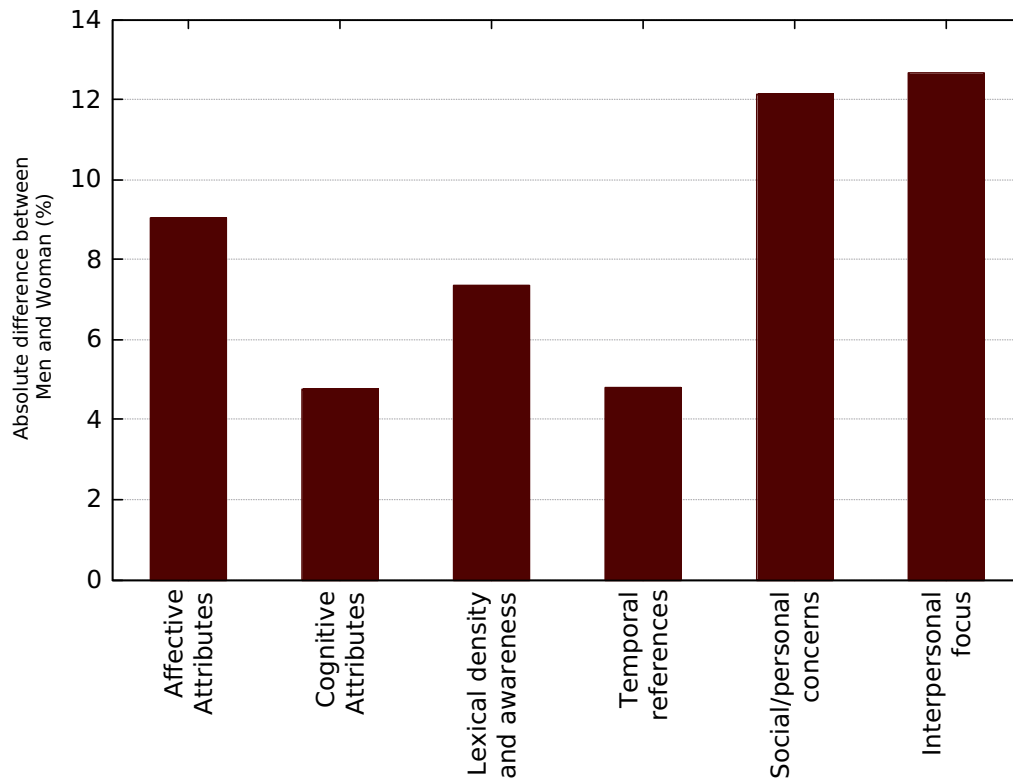
Figure 5.1: Mean absolute differences between male and female users per various categories of linguistic measures

In the race domain, the analysis of the linguistic difference for each race was performed in the same way as gender, but considering the other two races combined. Figure 5.2 shows the mean absolute differences between white and black/asians combined. As we can see, there is a stronger difference in affective attributes, which comprises the expression of anger, anxiety, sadness, and swear. Other linguistic aspects such as social/personal concerns and interpersonal focus showed to be relevant when comparing the writing of white users against the black and asian group.

Respectively, the linguistic difference among black users was compared against white and asian users combined. Again, affective attributes are the linguistic group with the features that most differ from one ethnicity to the others.

When it comes to comparing the asian linguistic to that in white and black users, some group of features that did not present higher absolute differences when comparing
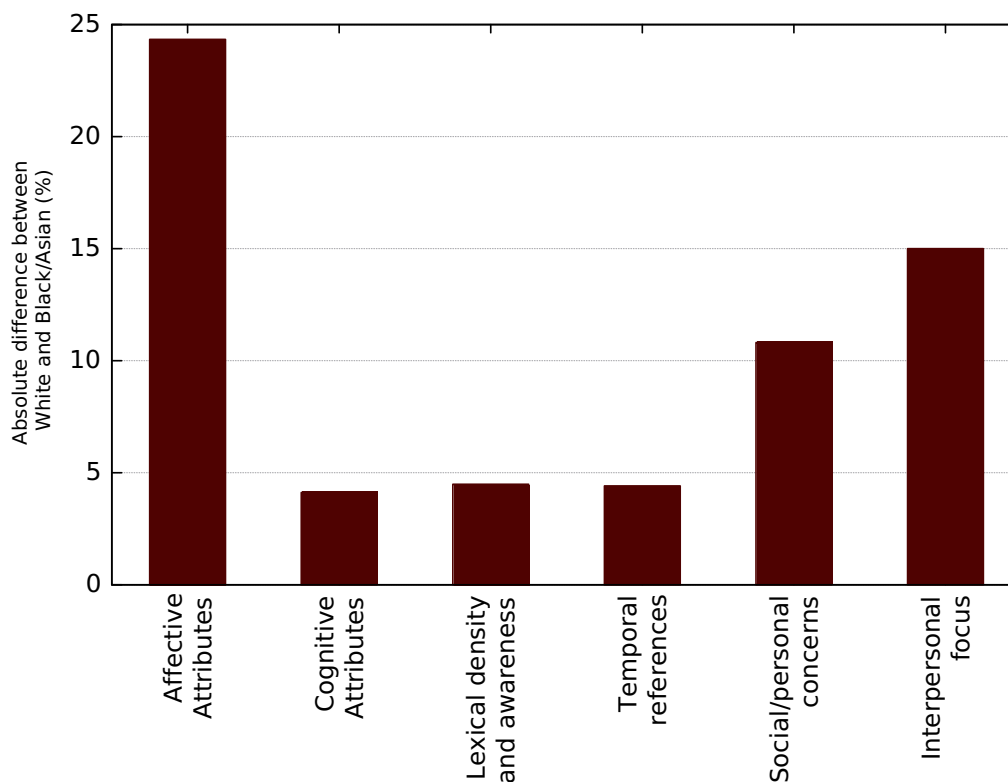
Figure 5.2: Mean absolute differences between white and black/asian users combined per various categories of linguistic measures

black and white groups, now tend to be higher such as lexical density and awareness and temporal references, which reveal some differences reflected by such different cultures especially in their way of writing.

Additionally, we correlate the produced linguistic features with gender using Wilcoxon rank sum significance tests. As Table 5.1 presents, females tend to use anxiety ($z = -74.534$) and sadness ($z = -74.394$) terms and phrases. On the other hand, males express anger ($z = 4.733$) in their tweets.

In terms of cognitive attributes, females are more likely to write phrases that express cognition and perception. From this group of features, two stand out: certainty ($z = -60.593$) and feel ($z = -70.766$) showing how females express more confidence and feelings in their writing.

In lexical density and awareness, we can see that females make more use of verbs ($z = -45.808$), auxiliary verbs ($z = -46.441$), conjunctions ($z = -72.098$), and adverbs ($z = -66.915$), while males use more articles ($z = 77.303$) and prepositions ($z = 32.596$).

The temporal references attributes are more present in the females writing, as we can see from the values for present tense ($z = -62.110$) and future tense ($z = -15.118$)
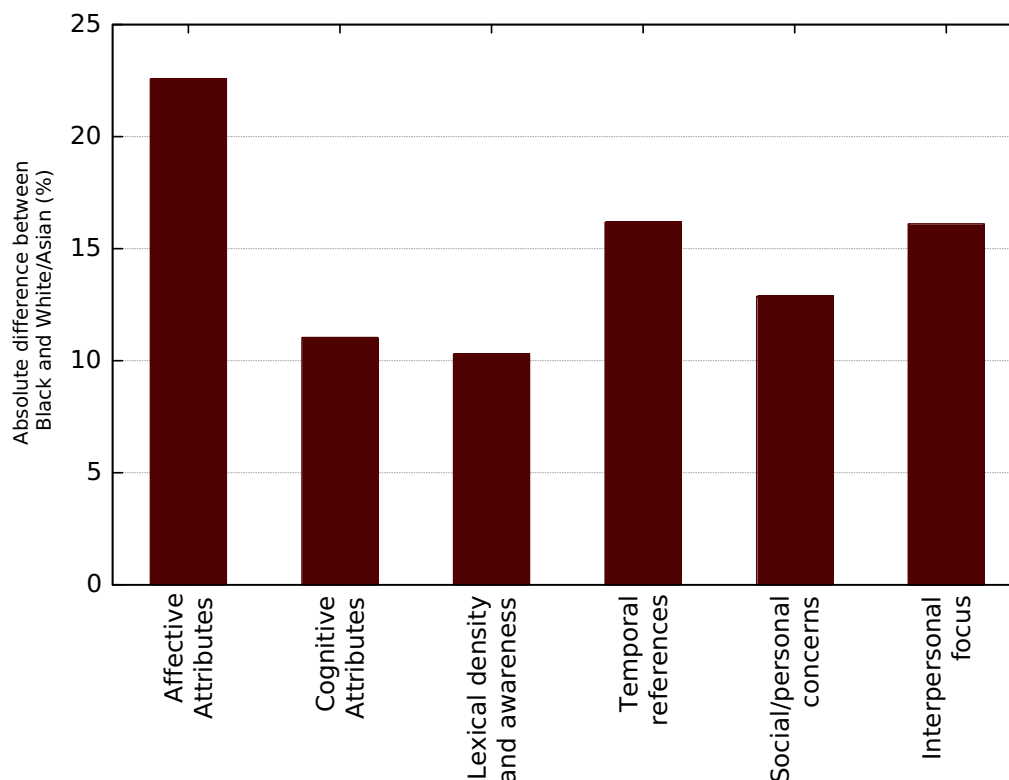
Figure 5.3: Mean absolute differences between black and white/asian users combined per various categories of linguistic measures

From social/personal Concerns perspective there is a clear trend on the usage of these features by females more than by males. Among the most noticeable values shown in Table 5.1, are family ($z = -93.252$), bio ($z = -102.681$). Also, the predominance of features like friends, social, health, and body show that females express more social and personal concerns in their writing than males. The only feature in this group that is more present in males' writing is achievement ($z = 65.265$)

Noticeably, females also have a higher tendency to write in the first person singular ($z = -97.329$) and in the second person ($z = -88.482$) than males, while there is a slight trend towards males using the first person plural in detriment of females ($z = 4.309$).

Also, from the race perspective, the difference of values between each race shows some particularities in the way of writing for each race. In this analysis, one race is compared with the other two combined (e.g. white users are compared with blacks and asians). From affective attributes, it is possible to see that black users tend to express more anger ($z = 94.610$) and swear ($z = 107.344$) than white/asian.

From cognitive attributes, almost all features were more present in black users texts than in the other races, with higher values for certainty ($z = 62.239$), hear
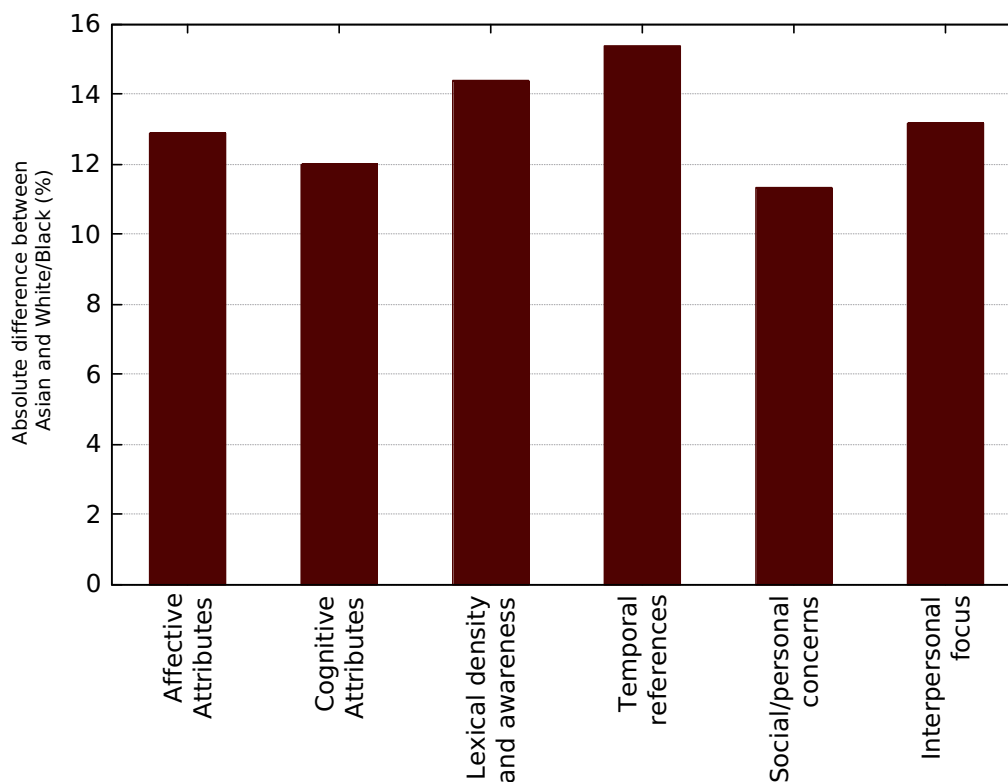
Figure 5.4: Mean absolute differences between asian and white/black users combined per various categories of linguistic measures

$(z = 62.137)$, and feel $(z = 63.963)$.

In terms of lexical density and awareness, black users have more presence in features like verbs, auxiliary verbs, conjunctions, and adverbs, while prepositions are more present among white users.

In regards to social/personal concerns, there is a higher presence of black people, noticeably in family $(z = 86.721)$, social $(z = 90.830)$, religion $(z = 85.163)$, and body $(z = 86.903)$.

The interpersonal focus feature set reveals that there is a predominance in the use of first person plural for white $(z = 77.425)$ while first person singular $(z = 63.492)$, second person $(z = 95.495)$ and third person $(z = 87.717)$ are more prominent in the black group.

Table 5.3 and Table 5.4 present the ranking difference for the 20 most common phrases for gender and races respectively. To find these differences, we randomly selected $1,000$ users from each group (male, female, asian, black, and white). Their tweets were used to create ngrams for each group. With this subset of our dataset, we extracted the top 100 phrases for each demographic group and the top 20 are shown in these Tables.

As we can see in Table 5.3 phrases expressing negation are in the top 4-positions for both males and females. It is also clear to see that females are more into signs than males since phrases with this kind of content present higher differences in the gender ranking.

Due to the informal nature of Twitter, the top phrases also reveal that it is common the usage of slangs like "do n't", "ca n't" and "wan na" for both genders.

When analyzing the ranking of race top phrases in Table 5.4, the trend of using negation phrases also happens here. Phrases containing expressions like "i don't", "i can't" and "i'm not" appear in the top positions for all the racial groups. Another interesting result is the position of the expression "i love you" in the writing of different races. White and asian users seem to be more likely to tweet contents with this expression than black users. Also, the expression "i want to" appears more often in the writing of white and asian users than in the blacks. These tables show differences regarding the way of writing of each demographic group and reveal interesting characteristics about the difference from one to another.

5.2   Differences in Topic Interests

Males and females may have differences in preferences and interests in digest information. In order to understand which topic is preferable to females than males, we analyze the differences in the topic interest of users in our dataset. The Figure 5.5 shows the gender distribution for the 20-top topics that we extracted, with log-ratio of perceived male to female. It shows the topic interest for users based on gender in our dataset. On the right side, we see topics related to males' interests while on the left side we see the topics that females are more interested than males. The 3-top topics for males are sports, organizations, and technology. In other words, males tend to interest more in these topics than females. However, females interest more for life, actors, and movie than males. More specifically, the gender difference between topics varies among males and females.

In a similar way, we present the race distribution for the 20-top topics of asian, black, and white users in Figure 5.6. In order to show results regarding race, for this specific analysis, we have normalized the dataset by the number of black users once they are the minority amount of users in our dataset, as shown in Table 3.3. Therefore, we have randomly selected $45,406$ users for each race to study their topic interests. Users from different races may also vary in interests and preferences. Figure 5.6-a shows that white users have more interest in politics, writers, and organizations than asians.
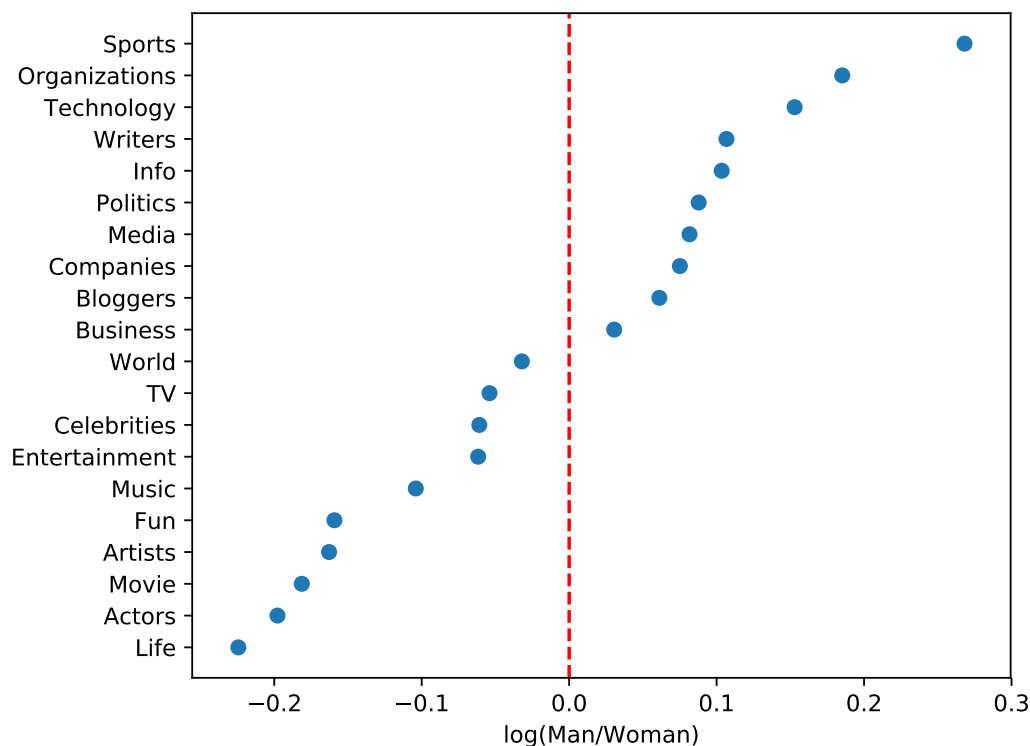
Figure 5.5: Gender interests: dots represent the gender interests for the 20-top popular topics.

However, asians prefer more artists, actors, and music topics than whites. Figure 5.6-b compares the differences in topic interests for white and blacks. We see that white users are interested in technology, movie, and politics more than blacks. Nonetheless, blacks prefer more artists, life, and music topics. Finally, from Figure 5.6-c, we see that asians have more interest for movie, companies, and technology topics than blacks. On the other hand, blacks prefer more business, sports, and organizations than asians.

5.3 Concluding Remarks

In this chapter, we presented the linguistic differences between gender and race. Further, we also analyzed the differences in terms of interests. We showed that interpersonal focus, which contemplates features like family, friends, health, religion, body, achievement, home, and sexual is the most prominent linguistic differences among males and females. There is a stronger difference in affective attributes for white and black users in comparison to other users, which comprises the expression of anger, anxiety, sadness, and swear. The differences in lexical density and awareness and temporal references attributes are the most significant between asians and other groups. This
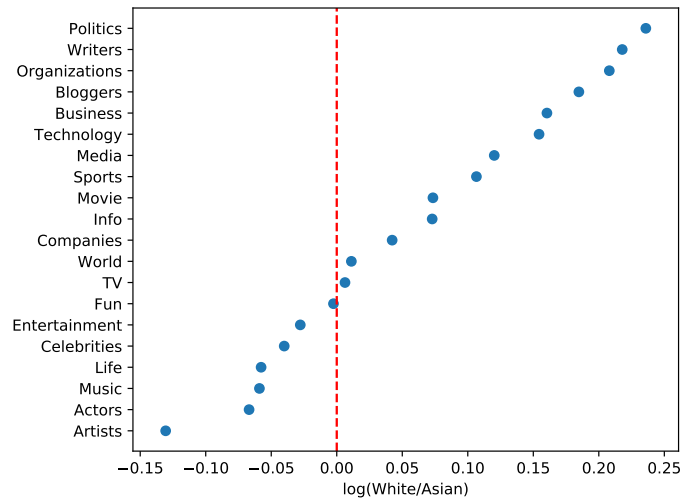
(a) White vs Asian
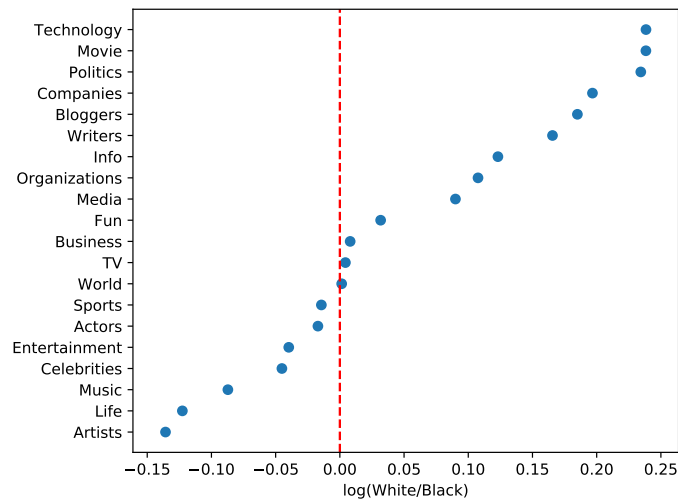
(b) White vs Black
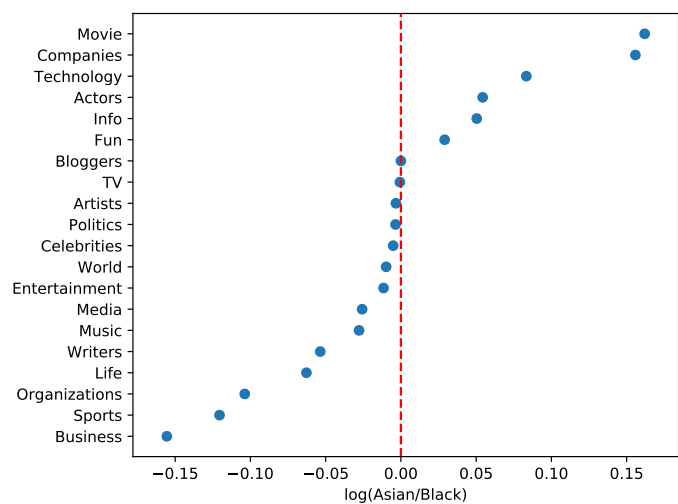
(c) Asian vs Black

Figure 5.6: Race interests: dots represent the race interests of (a) white against asians, (b) white against blacks, and (c) asian against blacks for the 20-top popular topics. The dataset is normalized by the number of blacks as shown in Table 3.3.

reveals some differences reflected by these different ethnicity especially in their way of writing.

Other interesting findings are that females tend to use anxiety and sadness terms while males express more with anger in their tweets. Females are more likely to write phrases that express cognition, perception, confidence and feelings. Additionally, females make more use of verbs, auxiliary verbs, conjunctions, and adverbs while males use more articles and prepositions. The temporal references and social/personal concerns attributes are more present in the females. Also, they have a higher tendency to write in the first person singular and in the second person while males use the first person plural and express more concern of achievement. In a similar analysis, black users tend to express more anger, swear, and cognitive attributes than white/asian. They also have more presence in features like verbs, auxiliary verbs, conjunctions, and adverbs. Blacks, tend more to use terms related to family, social, religion, and body. On the other hand, prepositions and the use of first person plural are more present among white users while the third person is more prominent for blacks.

Then, in terms of interests, we showed that the 3-top topics for males are sports, organizations, and technology. On the other hand, females have more interest for life, actors, and movie. Regarding race, we showed that white users have more interest in politics, writers, and organizations than asians. Asians prefer to write about artists, actors, and music topics more than whites. White users are more interested in technology, movie, and politics than blacks while blacks prefer more artists, life, and music topics. Asians have more interest for movie, companies, and technology topics than blacks. Finally, blacks prefer more business, sports, and organizations than asians.

Finally, phrases expressing negation are in the top positions for both males and females. It is also clear to see that females are more into zodiac signs than males since phrases with this kind of content present higher differences in the gender ranking. Further, white and asian users seem to be more likely to tweet contents "I love you" and "I want to" than blacks.

In the next chapter we discuss how different demographic groups are connected and interact with each other. For doing so, we dispose of the followers and friends of all users from our dataset. Our ultimate goal is to examine the proportion of connections and interactions among each demographic group.

Table 5.1: Differences between tweets from male and female users based on linguistic measures. $\mu(male)$ and $\mu(female)$ are the median value of the feature for male and female respectively. Statistical significance is count based on *Wilcoxon rank* sum tests. Linguistic features are extremely significant except from *hear* (*:=significant) based on *p*-values.

	$\mu(male)$	$\mu(female)$	z
Affective attributes			
anger	0.0055	0.0056	4.733
anxiety	0.0016	0.0019	-74.534
sadness	0.0029	0.0034	-74.394
swear	0.0023	0.0026	-7.411
Cognitive attributes			
Cognition			
causation	0.0101	0.0104	-18.627
certainty	0.0101	0.0111	-60.593
tentativeness	0.0136	0.0141	-14.641
Perception			
see	0.00957	0.0099	-24.538
hear	0.0055	0.0056	-0.033*
feel	0.0035	0.0041	-70.766
percepts	0.0207	0.0218	-41.373
insight	0.0115	0.0125	-46.806
relative	0.1014	0.0999	18.026
Lexical Density and Awareness			
verbs	0.1103	0.1170	-45.808
auxiliary verbs	0.0539	0.0583	-46.441
articles	0.0370	0.0340	77.303
prepositions	0.0843	0.0817	32.596
conjunctions	0.0279	0.0314	-72.098
adverbs	0.0317	0.0355	-66.915
Temporal references			
present tense	0.0802	0.0871	-62.110
future tense	0.0103	0.0106	-15.118
Social/Personal Concerns			
family	0.0026	0.0034	-93.252
friends	0.0028	0.0033	-66.168
social	0.0938	0.1021	-77.896
health	0.0037	0.0044	-76.446
religion	0.0024	0.0025	-26.485
bio	0.0157	0.0203	-102.681
body	0.0045	0.0056	-58.386
achievement	0.0116	0.0105	65.265
home	0.0022	0.0026	-74.049
sexual	0.0011	0.0012	-18.691
death	0.0014	0.0013	29.463
Interpersonal focus			
1st p. singular	0.0245	0.0340	-97.329
1st p. plural	0.0046	0.0045	4.309
2nd p.	0.0160	0.0198	-88.482
3rd p.	0.0030	0.0031	-3.371***

Table 5.2: Differences between tweets from white, black, and asian users based on linguistic measures. $\mu(White)$, $\mu(Black)$ and $\mu(Black)$ is the median value of features for each demographic group respectively. Statistical significance is count based on Wilcoxon rank sum tests. The p-values are extremely significant for all linguistic features. We test the correlation of each unique demographic group with the others.

	$\mu(White)$	$\mu(Black)$	$\mu(Asian)$	$z_{W/B-A}$	$z_{B/W-A}$	$z_{A/W-B}$
Affective attributes						
anger	0.0051	0.0081	0.0056	-67.261	94.610	-5.236
anxiety	0.0017	0.0019	0.0016	-0.696	33.789	-30.517
sadness	0.0031	0.0034	0.0032	-20.814	28.205	-0.625
swear	0.0021	0.0064	0.0027	-90.375	107.344	11.329
Cognitive attributes						
Cognition						
causation	0.0104	0.0105	0.0096	29.931	19.465	-54.832
certainty	0.0105	0.0116	0.0101	-19.404	62.239	-33.955
tentativeness	0.0138	0.0152	0.0130	-8.958	55.174	-40.226
Perception						
see	0.0098	0.0098	0.0095	18.756	6.970	-29.506
hear	0.0055	0.0062	0.0054	-26.349	62.137	-25.331
feel	0.0037	0.0044	0.0039	-44.180	63.963	-5.128
percepts	0.0212	0.0223	0.0210	-14.067	43.711	-23.308
insight	0.0122	0.0128	0.0112	11.133	40.420	-51.201
relative	0.1020	0.1012	0.0936	50.614	15.841	-76.870
Lexical Density and Awareness						
verbs	0.1125	0.1222	0.1082	-16.435	64.214	-39.436
auxiliary verbs	0.0554	0.0612	0.0529	-12.202	58.285	-39.130
articles	0.0366	0.0339	0.0314	96.532	-26.056	-94.363
prepositions	0.0851	0.0817	0.0743	77.024	1.032	-95.556
conjunctions	0.0291	0.0319	0.0286	-11.852	43.571	-25.898
adverbs	0.0329	0.0363	0.0325	-17.239	48.159	-23.542
Temporal references						
present tense	0.0825	0.0912	0.0798	-21.972	69.126	-37.196
future tense	0.0103	0.0119	0.0099	-28.333	79.181	-38.719
Social/Personal Concerns						
family	0.0029	0.0040	0.0032	-74.318	86.721	10.755
friend	0.0031	0.0033	0.0033	-26.248	25.332	8.717
social	0.0956	0.1101	0.0971	-60.389	90.830	-10.166
health	0.0040	0.0044	0.0039	-9.579	45.973	-30.920
religion	0.0024	0.0031	0.0024	-53.672	85.163	-13.154
bio	0.0176	0.0204	0.0179	-32.215	53.914	-10.492
body	0.0048	0.0067	0.0052	-62.906	86.903	-3.428
achievement	0.0114	0.0109	0.0097	69.227	-1.632	-83.506
home	0.0025	0.0024	0.0022	50.362	-4.554	-57.624
sexual	0.0011	0.0019	0.0012	-51.768	71.799	-3.084
death	0.0014	0.0015	0.0013	4.356	31.454	-34.554
Interpersonal focus						
1st p. singular	0.0268	0.0355	0.0296	-51.874	63.492	4.760
1st p. plural	0.0048	0.0042	0.0039	77.425	-28.107	-68.994
2nd p.	0.0169	0.0227	0.0177	-63.930	95.495	-10.148
3rd p.	0.0030	0.0039	0.0028	-36.070	87.717	-37.143

Table 5.3: Ranking differences of gender top phrases. We use *ne* for no existing phrases in a group.

	Rank(female)	Rank(male)	Diff(F-M)
i do n't	1	1	0
i ca n't	2	2	0
you do n't	3	3	0
i 'm not	4	4	0
ca n't wait	5	8	3
i 'm so	6	19	13
i love you	7	15	8
do n't know	8	11	3
i want to	9	24	15
more for virgo	10	55	45
more for cancer	11	29	18
i wan na	12	28	16
! i 'm	13	25	12
you ca n't	14	16	2
more for libra	15	39	24
it 's a	16	10	6
and i 'm	17	33	16
more for pisces	18	ne	-
i need to	19	34	15
do n't have	20	27	7

Table 5.4: Ranking differences of race top phrases. We use *ne* for no existing phrases in a group.

	Rank(White)	Rank(Black)	Rank(Asian)	Diff(W-B)	Diff(W-A)	Diff(B-A)
i do n't	1	1	1	0	0	0
i ca n't	2	2	2	0	0	0
ca n't wait	3	18	7	15	4	11
you do n't	4	4	3	0	1	1
i 'm not	5	8	6	3	1	2
i love you	6	33	4	27	2	29
i 'm so	7	16	6	9	1	10
do n't know	8	19	11	11	3	8
it 's a	9	26	16	17	7	10
one of the	10	48	20	38	10	28
i want to	11	47	10	36	1	37
! i 'm	12	46	29	34	17	17
if you 're	13	28	19	15	6	9
thank you for	14	126	28	112	14	98
it 's not	15	34	32	19	17	2
and i 'm	16	58	21	42	5	37
you ca n't	17	17	17	0	0	0
i 'm at	18	53	26	35	8	27
n't wait to	19	100	51	81	32	49
i liked a	20	7	ne	13	-	-

Chapter 6

Demographic Group Interconnections

Next, we analyze how demographic groups are connected with each other. Our ultimate goal is to examine the proportion of connections and interactions among each demographic group. Thus, for this analysis we used the dataset of $428,697$ users related to friends, followers and interactions discussed previously.

In order to represent connection probabilities among demographic groups we built a probabilistic graph and we compute the probability of a connection/interaction for every pair of users. To gain a holistic view, we aggregate all pairs of users making the nodes to represent demographic groups and each direct edge to represent the probability of a relationship happened, like friendship or interaction. The sum of all outgoing probabilities for each demographic group is 1.0.

6.1 Gender and its Interconnections

We begin by analyzing interconnection among groups of users separated by gender. We analyze the probability of male and female users to be friends with (i.e. to follow) other male and female users. In our analysis, we note in Figure 6.1 that females are equally likely to follow males and females in 50%, while males tend to follow 56% of males versus 44% of females. If we take the numbers in Table 3.3 as a proxy for the expected population of males and females in a randomly created list of friends, we note that even the equally distributed list of friends of females is actually disproportional as the expected percentage of males and females are 48.12% and 51.88%, respectively. This means that males tend to follow 16.37% more other males than expected, whereas females tend to follow 3.9% more males than expected. This suggests that both, males

and females connections, take their part of the responsibility on these gender inequalities, but it also shows that males tend to take the larger part of the responsibility.
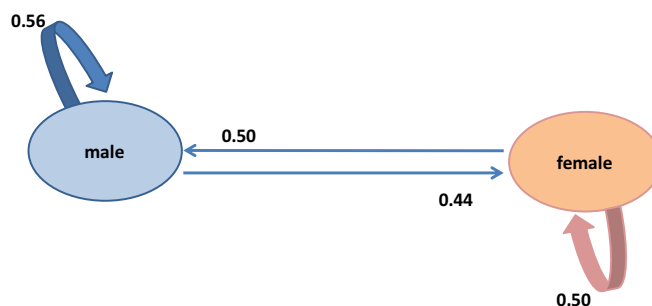
Figure 6.1: Probability of friendship for groups of users separated by gender
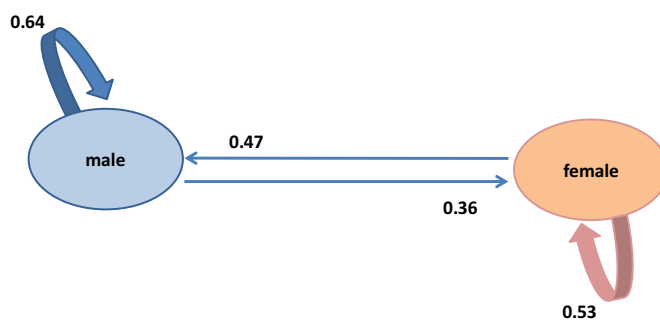
Figure 6.2: Probability of interactions for groups of users separated by gender

In terms of interactions, similar observations can be made. However, as we can see in Figure 6.2, the proportion of males and females retweet/mentioned by females is quite close to the expected population (53% of females against 47% for males). For the group of males, we noted a much higher disparity in terms of interactions in comparison to friendship. Males tend to interact with 64% of other males and only 36% of females.

This result quantifies the perceived inequality in Twitter. These observations are in line with perceived inequalities in the offline world, including in academic job hiring, earnings, funding and academic rating Sugimoto et al. [2013, 2015] as well as income differences between males and females in the population of high earners Merluzzi and Dobrev [2015].

6.2   Race and its Interconnections

Next, we turn our attention to the analysis of interconnection among groups of users separated by race. Figure 6.3 shows these interconnections in terms of friendship and interactions, respectively.

(a) Friendship
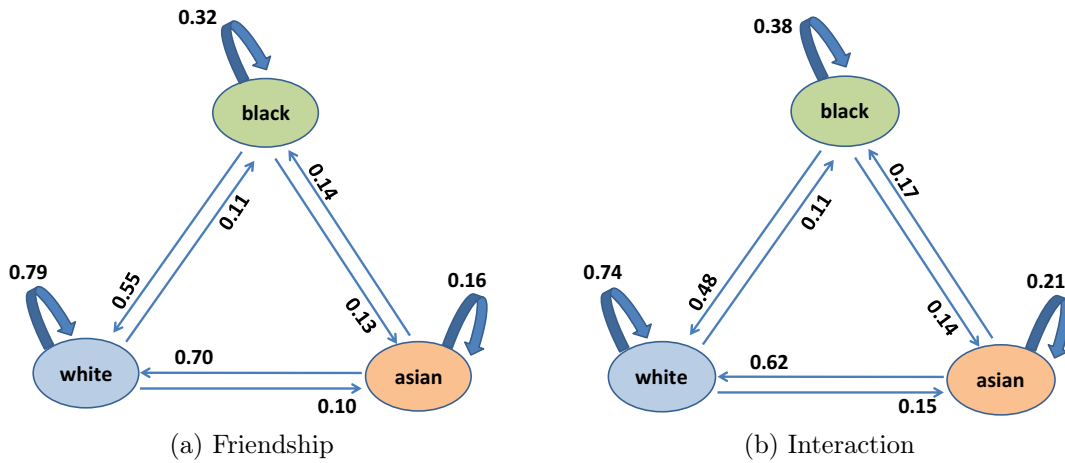
(b) Interaction

Figure 6.3: Probability of friendship and interaction for groups of users separated by race

We start by analyzing the self-loops on each node. We note that white users tend to follow about 79% of other white users, black users tend to follow 32% of black users and asians tend to follow 16% of asians. The key for analyzing these figures is to recall the reference distribution of users in our dataset according to their race, in which whites represent 67.97%, black accounts to 14.91%, and asian accounts to 17.12%. Thus, we can note that, in comparison to the expected distribution of users, whites tend to follow 16.22% more whites than expect, blacks tend to follow impressive 2.14 times more blacks than expect, and asians end up following less (6.54%) asians than expected. In other words, the presence of homophily was not clear for the case of asians.

Similar observations can be made for interactions, the only exception is among asians, where the fraction of interactions is higher than the reference distribution. Although asians tend to follow more white users (even more than the reference distribution), the interactions with white users are below the reference fraction. Overall, whites and blacks showed quite high numbers of endogenous connections and interactions as both groups seem to avoid following a different race. The stronger mutual interconnection is between black users.

6.3   Demography of Interconnections

Finally, we characterize the interconnection of users grouped by race and gender. Figure 6.4 shows these interconnections in terms of friendship and interactions, respectively. However, given the high number of edges in these figures, instead of providing

the transition probabilities, we computed their relative increase or decrease in comparison to the expected demographic population.

In addition, to reinforce previous findings from the analysis of gender and race separately presented before, these results show interesting trends. First, we note that white males tend to receive only positive edges from white and asian groups, meaning that only black users (both males and females) tend to have a fraction of white males users as friends that are relatively lower than the expected demographic distribution in Twitter. This suggests that race plays a more important role for connections with white males. In terms of interactions, similar observations hold, except that asian females tend to have a relatively smaller number of interactions with white males. Interestingly, white females appear over-represented only in the list of friends of asian females.

Another interesting pattern is that the highest values for self-loops are for black males (130%) and black females (86%). These numbers are higher for interactions, 232% and 164%, respectively. The interconnections between these two groups are much larger than expected. In contrast, the lower values for self-loops are for asians, with negative values for friendship, but positive for interactions, as discussed before. Interestingly, all incoming links to the two asian groups are negative, meaning that the other two demographic groups tend to have an under-represented proportion of asians as friends and in terms of interactions.

6.4   Concluding Remarks

In this chapter, we analyzed the interconnection among groups of users separated by gender and race. We analyzed the probability of males and females users to be friends with (i.e. to follow) other males and females users. We found that males and females connections take their part of the responsibility on gender inequality. However, males tend to take the larger part of the responsibility. Further, race plays a more important role for connections with white males. In terms of interactions, similar observations hold, except that asian females tend to have a relatively smaller number of interactions with white males. Our analysis of connections and interactions among these groups explain part of causes of such inequalities and offer hints for the promotion of equality in the online space. The next chapter present the system design of the *Who Makes Trends?* Web-based service in order to make the demographic biases in Twitter trends more transparent.
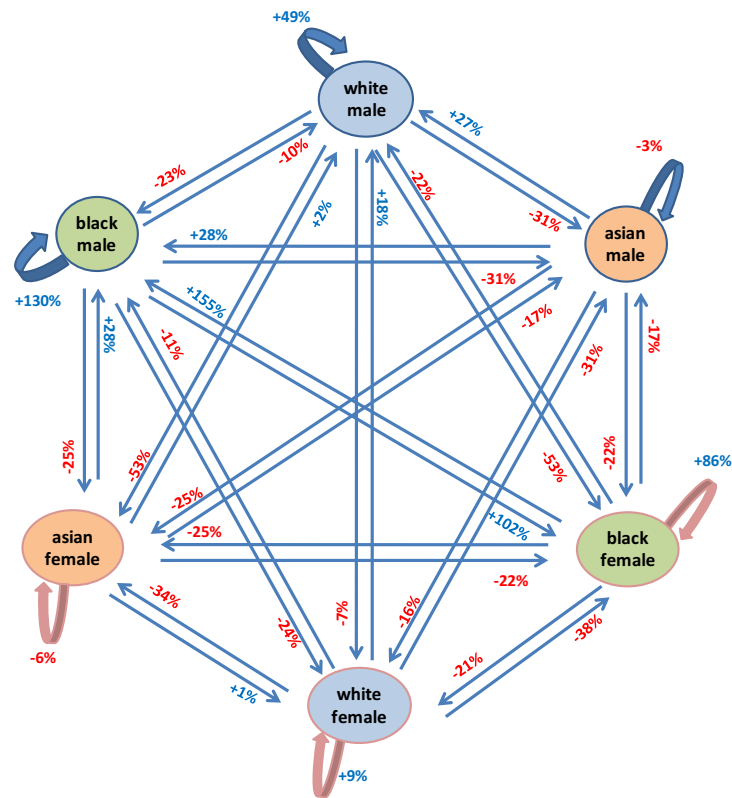
(a) Friendship
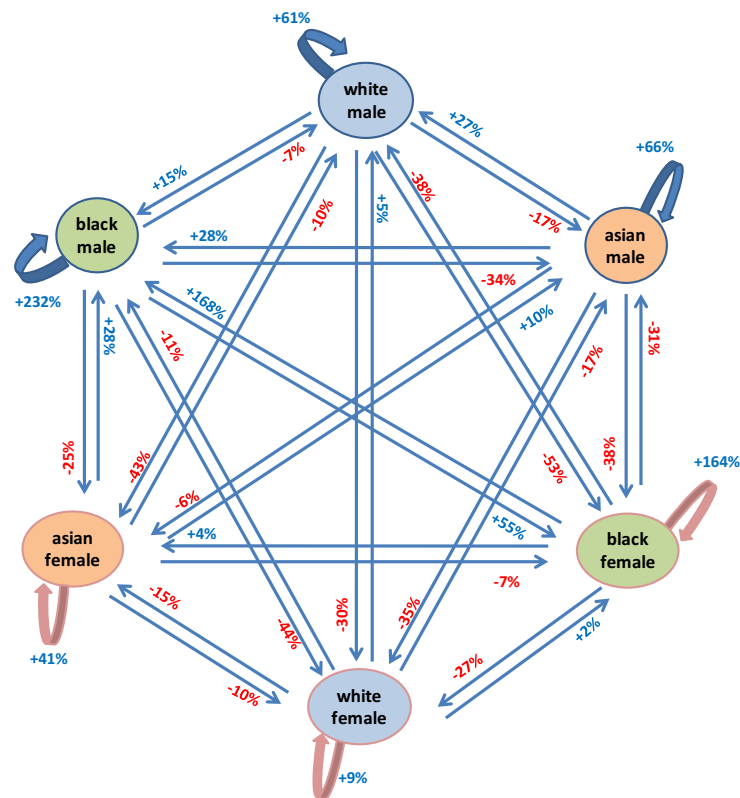
(b) Interaction

Figure 6.4: Relative probability of friendship and interaction for groups of users separated by race and gender. The value declares the relative increase (blue color) or decrease (red color) in comparison to the expected demographic population.

Chapter 7

Leveraging Demographic Aspects to Design Transparent Systems

Demographics aspects have a valuable power to provide transparency of algorithms and systems that use user-generated data. In this dissertation, we focus on providing transparency on the Twitter trending topics in order to show how powerful are demographic aspects for providing transparency. We believe this strategy can also be applied in different systems such as Google Suggestions. The ability to understand how Google suggestions algorithm works, by using demographic aspects, would allow us to analyze which demographic groups have created these suggestions and would help us to study stereotype, reappropriation, and many other concepts.

Increasingly, researchers and governments are recognizing the importance of making algorithms transparent. The White House recently released a report that concludes that the practitioners must ensure that AI-enabled systems are open, transparent, and understandable Felton and Lyons [2016]. Hence, there is a concern in promoting more transparent data to users.

Users of social media sites like Facebook and Twitter rely on crowdsourced content recommendation systems (such as Trending Topics) to retrieve important and useful information. Once the contents are chosen for the recommendation, they reach a large population, actually giving the initial users of the contents an opportunity to promote their messages to the wider public. Consequently, it is important to understand the demographics of users who make a content worthy of recommendation, and explore whether they are representative of the media site's overall population. Therefore, to make demographic bias more transparent in Twitter we developed and deployed the *Who Makes Trends?*[1] Chakraborty et al. [2017] Web-based service. It infers the

---

[1] http://twitter-app.mpi-sws.org/who-makes-trends/

demographic of Twitter trending topics in U.S.

7.1 Who Makes Trends?

To make the demographic biases in Twitter trends more transparent, we developed and
deployed a real-time Web-based service *Who Makes Trends?* at `http://twitter-app.`
`mpi-sws.org/who-makes-trends`. It shows information of the Trend Promoters and
Trend Adopters. Trend Promoters is the distribution (or combination) of demographic
groups (such as middle-aged white males, young asian females, adolescent black males)
in the crowd promoting (or posting about) a topic before the topic becomes trending
in Twitter. On the other hand, Trend Adopters is the distribution (or combination)
of demographic groups in the crowd adopting a topic after the topic becomes trending
in Twitter. Here, we are only considering US-based Twitter users whose tweets on the
trends appear in the 1% random sample we obtained from Twitter.

Figure 7.1 shows the main page (index page) of the service. It contains three
sections: (i) the search trends and sample trends with high demographic bias; (ii) how
it works; and (iii) who are we?. The first section allows users to search trends by
text (see Figure 7.2) or even by date (see Figure 7.3). It also contains some trending
hashtags manually selected with high gender bias, high racial bias, and high age bias
as an example of data transparency. The second basically points to the Chakraborty
et al. [2017] work and explains more details about gender, race, and age bias in Twitter
trending topics. Finally, the last section shows the people who worked on this project.

This section describes necessary procedures to develop this application. Basically,
it describes the data collection and methods to provide the distribution of Promoters
and Adopters Trends of Twitter trending topics. Detailed information is available in
Chakraborty et al. [2017] work.

7.1.1 Data collection

For this work, we gather and store, every day, the 1% random sample of all English
tweets posted in U.S, through the Twitter Stream API[2], since January 1st, 2017. The
filtering process is configurable via parameter setting in the Twitter Stream API. There-
fore, we use as a filter the U.S. bounding box to infer tweets from U.S. and also the
language as English. After this process, we are able to get all English tweets from U.S.
provided by the Twitter Stream API. We prefer to use the Stream Filter because it
returns more tweets from U.S. than the worldwide stream since 1% of the total tweets

---

[2]`https://dev.twitter.com/streaming/overview/request-parameters`

Figure 7.1: Main page of *Who Makes Trends?* web-based service. It allows users to search trends by text or by date. On this page, there are some manually selected hashtags which contain high gender bias, high racial bias, and high age bias as an example of data transparency.
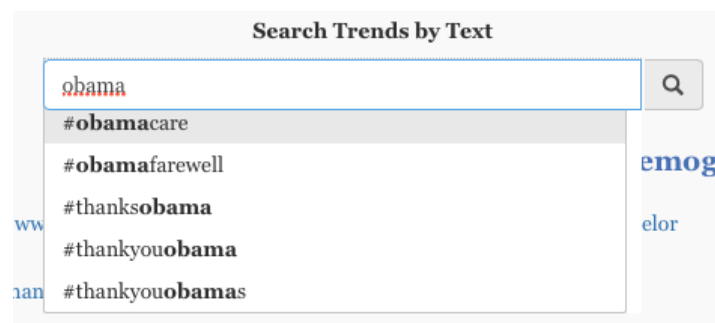
Figure 7.2: Search Trends by Text box with auto-complete functionality. It allows users to search for a specific query and it also suggests some trending hashtags based on the input provided by users.

posted in the U.S. is higher than the total 1% posted worldwide. Simultaneously, and along the same period, by querying the Twitter REST API[3] at every 5-minutes, we collected all topics which became trending in the US. It is important to highlight, once we gather data from U.S., we use the Eastern Standard Time (EST)[4] as a default time zone in order to control the day (12 am to 11:59:59 pm) we gather the data. Addi-

---

[3]http://dev.twitter.com/rest/reference/get/trends/place
[4]https://www.timeanddate.com/time/zones/est
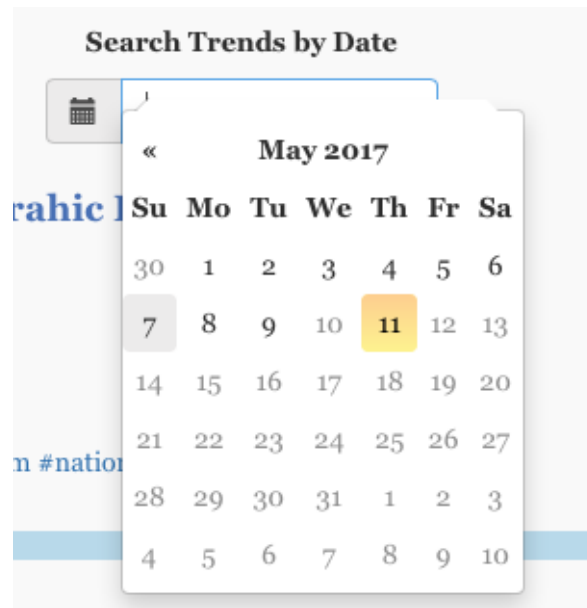
Figure 7.3: Search Trends By Date box which allows users to see which trends (we have data) appear on a specific day.

tionally, some hashtags may appear as trending in consecutive days. Hence, in order to reflect this restriction in our service, we decided to update the service every 2 days. Therefore, we process the data for the day D as being the data for the day D - 12 hours and also day D + 12 hours. This step is necessary because some hashtags may have been posted the day before (D - 1 day), became trending on day D and/or even become trending one day after (D + 1 day).

Once we have the daily Twitter data with all English tweets posted in U.S., the next step is to filter those tweets that contain at least one hashtag which appeared in the Twitter Trending Topic on that specific day. This step is necessary because the Web service shows demographic information of users who promote (post a tweet before it becomes trending) or adopt (post a tweet after it becomes trending) this trending hashtag. To do that, for every tweet in our daily dataset, we look at the tweet field *entities/hashtags* which provides all hashtags available in a tweet and we check if the hashtag appeared in the Twitter Trending Topic. Then, in order to show the demographic information regarding Trend Promoters and Trend Adopters we create three set of users: (i) who posted a tweet before it becomes trending; (ii) who posted a tweet after it becomes trending; and (iii) the union of all users we could infer the demographic information since January 1st. This is possible because each trending topic gathered from Twitter has the precise time it became trending. Also, we would like to show the expected demographic distribution for each trending in order to compare the Trend Promoters and Trend Adopters as being under- or over-represented with respect to the

Twitter users population for which we have demographic data.

Now, once we have the three set of users, (i) Trend Promoters; (ii) Trend Adopters; and (iii) Expected Distribution, similarly to the methodology in Section 3.2, we submit the profile picture Web links into the *Face++ API*. We also discard those users whose profile pictures do not have a recognizable face or have more than one recognizable face, according to Face++. Finally, the service is able to show the demographic distribution of Twitter Trending Topics.

7.1.2 Trending Topic Analysis

To compare the demographics of Twitter users with the demographics of the *offline* population, we collect the demographics of U.S. residents from the U.S. Census Bureau[5] [6] and present in Table 7.1. We see that some demographic groups appear more in Twitter compared to their share of U.S. population. For instance, the presence of asians in Twitter is about 3.06 times larger than in the overall U.S. population. Similarly, the adolescent and young people are much more in Twitter. However, mid-aged and old population have comparatively significantly less presence in Twitter. Our findings corroborate a recent survey on social media population conducted by Pew Research[7].

Table 7.1: Comparing the demographics of the population in U.S., and the demographics of U.S.-based Twitter users, whose tweets were included in the 1% random sample during January – May 2017, and whose demographic information could be inferred.

Baseline	Gender (%)		Race (%)			Age Group (%)			
	Male	Female	White	Black	Asian	Adolescent	Young	Mid-aged	Old
U.S. Population	49.20	50.80	72.40	12.60	4.80	13.60	26.70	33.20	13.50
Twitter Population	45.97	54.03	73.05	12.25	14.70	26.37	62.58	10.80	0.25

Table 7.2, shows the amount of promoters, adopters, and the total users with inferred demographic information and also the total of users we could not infer the demographics who used the specific trending hashtag from April 2nd to May 2nd, 2017, sorted in a descending order based on the date. As an interesting finding, some trending hashtags appear to be promoted by Twitter even when there are not so many users posting tweets about this topic. We believe it may happen due to some important and well-known events. Therefore, to give more visibility to those contents,

---

[5]https://www.census.gov/topics/population/age-and-sex.html
[6]http://census.gov/prod/cen2010/briefs/c2010br-02.pdf
[7]http://pewinternet.org/2016/11/11/social-media-update-2016

Twitter chooses them as trending. For example, the trending hashtags *#fyrefestival*, *#unicornfrappuccino*, and *#eastersunday*.

#fyrefestival[8,9] (Fyre Festival): is regarding the music festival *Fyre Festival* scheduled to happen in Bahamian island of Great Exuma. However, the festival was a disaster, grounds that were woefully lacking the promised amenities, replaced instead by dirt fields, soggy tents, and folding chairs. Hence, people got upset and started to post in Twitter what was going on in the event. The organizers had to postpone the festival.

#unicornfrappuccino[10,11] (Unicorn Frappuccino): It is a drink created by Starbucks in 2017. However, it got viral because users did not like it and started to post jokes about this new drink.

#eastersunday[12] (Easter Sunday): It celebrates the resurrection of Jesus. It is an event celebrated by Christian culture around the world which made it appears as trending in Twitter.

7.1.3   Disparate Demographics

While analyzing the demographics of the promoters of different trends, we observed that different trends are promoted by user-groups having highly disparate demographics. We classify these disparate demographics into three categories: (i) high gender bias; (ii) high racial bias; and (iii) high age bias.

High gender bias contains trending hashtags that are highly under- or over-expected for gender. As examples, we have the hashtags #footballmovies, #ufcphoenix, and #thebachelor. On the other hand, high racial bias contains trending hashtags that are highly under- or over-expected for race. As examples, we have the hashtags #thankyoutrump, #obamacare, and #neweditionbet. Finally, for the high age bias, we have the hashtags #dangerouswomantour, #presidentialtvshows, and #nationalloveyourpetday which are highly under- or over-expected for age. Table 7.3 shows the demographic distribution of promoters of different trending hashtags. We can see that trending hashtags like #thankyoutrump was promoted by mid-aged white users, similarly, #obamacare was promoted by mid-aged white males. On the other hand, trending hashtags like #neweditionbet were promoted by adolescent black females.

---

[8]http://www.fyrefestival.com
[9]http://www.nytimes.com/2017/04/28/arts/music/fyre-festival-ja-rule-bahamas.html
[10]http://en.wikipedia.org/wiki/Unicorn_Frappuccino
[11]https://www.cnet.com/news/starbucks-unicorn-frappuccino-comparisons-twitter-jokes/
[12]http://en.wikipedia.org/wiki/Easter

Table 7.2: Top hashtags during 31 days from April 2nd to May 2nd, 2017. They are sorted in a descending order based on the date. This Table shows the number of promoters, adopters, and the total users with inferred demographic information. Also, it shows the number of total users who we could not infer the demographic information. The hashtags are clickable links that point to their demographic distribution on *Who Makes Trends?* service.

Hashtag	Date	#Promoters With Demographic Inference	#Promoters Without Demographic Inference	#Adopters With Demographic Inference	#Adopters Without Demographic Inference	#Total With Demographic Inference	#Total Without Demographic Inference
#mayday2017	02-05-2017	609	532	162	165	736	660
#metgala	01-05-2017	1563	616	491	257	1988	830
#wwepayback	30-04-2017	862	660	88	70	912	695
#climatemarch	29-04-2017	637	582	155	142	753	694
#fyrefestival	28-04-2017	32	15	1363	925	1384	936
#nfldraft	27-04-2017	4846	3635	2318	1873	6183	4813
#wednesdaywisdom	26-04-2017	317	341	77	57	383	392
#dwts	25-04-2017	360	188	89	51	435	232
#mondaymotivation	24-04-2017	673	717	141	131	803	840
#sundayfunday	23-04-2017	678	613	153	112	812	712
#earthday	22-04-2017	2636	2719	3105	2618	5531	5123
#ripprince	21-04-2017	453	300	171	104	603	393
#happy420	20-04-2017	805	567	132	121	918	661
#bostonmarathon	19-04-2017	789	598	282	225	1005	778
#unicornfrappuccino	18-04-2017	36	13	1712	996	1737	1005
#cleveland	17-04-2017	709	442	406	355	1049	711
#eastersunday	16-04-2017	64	77	1570	1233	1616	1300
#aprilthegiraffe	15-04-2017	810	421	76	56	872	467
#goodfriday	14-04-2017	1674	1422	862	640	2452	1996
#stanleycup	13-04-2017	369	303	600	493	842	709
#bucciovertimechallenge	12-04-2017	171	244	770	992	867	1137
#nationalpetday	11-04-2017	2637	1887	1455	891	4056	2744
#nationalsiblingsday	10-04-2017	3828	1837	2512	1202	6296	3023
#sundayfunday	09-04-2017	775	585	181	130	939	705
#nationalbeerday	08-04-2017	1188	1581	368	333	1529	1876
#syria	07-04-2017	1263	856	654	472	1771	1217
#themasters	06-04-2017	420	452	3015	2513	3225	2771
#13reasonswhy	05-04-2017	280	87	985	363	1225	439
#nationalchampionship	04-04-2017	3376	2561	238	174	3533	2682
#finalfour	03-04-2017	4146	3448	851	618	4739	3887
#openingday	02-04-2017	1732	1342	4129	3437	5461	4506

Table 7.3: Demographics of promoters of Twitter trends. Demographic groups shown in bold blue are represented more (over-expected), and groups in red italics are represented less (under-expected) among the promoters. We consider differences of 5% as the threshold. The hashtags are clickable links that point to their demographic distribution on *Who Makes Trends?* service.

| Hashtag | Demographics of Promoters | | | | | | | |
| | Gender (%) | | Race (%) | | | Age Group (%) | | |
	Male	Female	White	Black	Asian	Adolescent	Young	Mid-aged
#footballmovies	**65.82**	*34.18*	**83.55**	*5.06*	11.39	*10.13*	**70.88**	**18.99**
#ufcphoenix	**77.03**	*22.97*	73.65	10.81	15.54	*16.89*	**71.62**	11.49
#thebachelor	*15.61*	**84.39**	**84.69**	*4.94*	10.37	29.82	64.94	*5.24*
#thankyoutrump	49.55	50.45	**81.98**	8.11	9.91	21.62	*54.96*	**22.52**
#obamacare	**58.11**	*41.89*	**83.78**	*6.76*	*9.46*	*13.51*	*51.26*	**32.43**
#neweditionbet	*40.66*	**59.34**	*28.27*	**58**	13.73	**33.51**	59.93	6.49
#dangerouswomantour	*36.67*	**63.33**	71.67	8.33	**20**	**43.33**	*50*	6.67
#presidentialtvshows	**68.31**	*31.69*	**80.33**	10.93	*8.74*	*8.20*	**72.67**	**18.58**
#nationalloveyourpetday	*28.49*	**71.51**	**80.27**	8.04	11.69	26.94	63.29	9.77

7.2   Concluding Remarks

In this chapter, we developed and deployed the *Who Makes Trends?* Web-based service in order to make the demographic biases in Twitter trends more transparent. To accomplish this design, we presented our methodology to build the system. Therefore, we described the data collection and trending topic analysis steps. We also observed that different trends are promoted by user-groups having highly disparate demographics. We classify these disparate demographics into three categories: (i) high gender bias; (ii) high racial bias; and (iii) high age bias. The *Who Makes Trends?* Web-based service is available at `http://twitter-app.mpi-sws.org/who-makes-trends/`. In the next chapter, we conclude this dissertation, discuss some limitations, and propose future research directions.

Chapter 8

Concluding Discussion and Future Work

Exploring demographic aspects is important both from the perspective of systems and from the sociological point of view. Hence, this dissertation characterizes demographic aspects of users in Twitter, identifying inequalities and asymmetries in gender and race. To do that, we gathered gender and race of Twitter users located in the United States using advanced image processing algorithms from Face++. We collected a large sample of $1,670,863$ users located in the United States with identified demographic information. We have also crawled a large sample of these users' tweets and friends to study connections and interactions among demographic groups of users.

Our analysis corroborates existing evidence about gender inequality in terms of visibility and introduces race as a significant demographic factor, which reveals hiding prejudices between demographic groups. We also show that the Twitter glass ceiling effect typically applied to females. Also, it occurs in Twitter for blacks and asian males.

We also attempt to characterize the differences in the way of writing for each group pointing the most important linguistic aspects for a specific gender and race. We were able to identify for each demographic group which affective attributes were more prominent in their writing. In the same way, features based on cognitive attributes, temporal references, social and personal concerns, and interpersonal focus showed to have different weights throughout different demographic domains. By analyzing topic interests, we found that each demographic group tends to have its own preferences for the information they share. For instance, we found that males are more into sports, organizations, and more interest in technology while females have more interest in topics related to life, actors, and movie. In the same way, users from different races are also likely to have different interests and preferences. white users are more interested

in politics, writers, and organizations when compared to asians, and technology, movie, and politics when compared to black users. On the other hand, black users are more into artists, life, and music topics. When we look into asians, they are more interested in artists, actors, and music than whites and tend to have more interest for movie, companies, and technology when compared to blacks. Another interesting conclusion is based on the most common phrases encountered on each group and their ranking position when compared to different demographic groups. The analysis of these most common phrases led us to conclude that phrases expressing negation figure as one of the most frequent for all domains. Also, the usage of slang, which is common in an environment like Twitter, appears in these frequent phrases too. One interesting point is that, when we compare the difference between the groups, we find interesting trends, like the more prominent interest in zodiac signs among females than among males.

Our effort is the first to quantify to what extent one demographic group follows and interact with each other and the extent to which these connections and interactions reflect in inequalities in Twitter. The connections and interactions among these groups explain part of causes of such inequalities and offer hints for the promotion of inequalities in the online space.

As potential limitations, we have the accuracy of Face++ and our approach to identify users located in the United States. The accuracy of Face++ inferences is an obvious concern in our effort because we rely on this strategy to infer the demographic aspects of Twitter users. Also, our approach to identify users located in the U.S. may bring together some users located in the same U.S. time zone, but from different countries. We believe that these users might represent a small fraction of the users.

Regarding the system design, demographics aspects have a valuable power to provide transparency of algorithms and systems that use user-generated data. Therefore, in this dissertation, we developed and deployed the *Who Makes Trends?* Web-based service in order to make the demographic biases in Twitter trends more transparent and show how powerful are demographic aspects in providing transparency. We believe this strategy can also be applied in different systems such as Google Suggestions and it would help us study stereotype, reappropriation, and many other social concepts.

Finally, there are some future directions we would like to pursue next. First, we plan to explore age as a demographic factor in social media. Second, we plan to examine how diffusion of news happens when it is propagated through demographic group interconnections and communications. Second, we believe that the increasing availability of information about demographics would help the development of systems that promote more diversity and less inequality to users. Then, we want to study the correlation of linguistic differences with other demographic factors e.g age. We plan to

use our extracted linguistic characteristics as a feature vector for prediction of gender and race. Moreover, we plan to examine the speed of tweets that are propagated through a specific demographic group.

Bibliography

Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. (2009). Diversifying search results. In *ACM WSDM*.

An, J. and Weber, I. (2016). #greysanatomy vs. #yankees: Demographics and hashtag use on twitte. In *Proceedings of the 10th International AAAI Conference on Web and Social Media*.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May*, 23.

Araújo, C. S., Meira Jr, W., and Almeida, V. (2016). Identifying stereotypes in the online perception of physical attractiveness. In *International Conference on Social Informatics*, pages 419--437. Springer.

Ayres, I., Banaji, M., and Jolls, C. (2015). Race effects on ebay. *The RAND Journal of Economics*, 46(4):891--917.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104:671.

Barr, A. (2015). Google mistakenly tags black people as 'gorillas,'showing limits of algorithms. *The Wall Street Journal*.

Bertrand, M. and Mullainathan, S. (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *The American Economic Review*, 94(4):991--1013.

Bhattacharya, P., Zafar, M. B., Ganguly, N., Ghosh, S., and Gummadi, K. P. (2014). Inferring user interests in the twitter social network. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 357--360, New York, NY, USA. ACM.

Blevins, C. and Mullen, L. (2015). Jane, john... leslie? a historical method for algorithmic gender prediction. *Digital Humanities Quarterly*, 9(3).

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1--8.

Bonilla-Silva, E. (2006). *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States.* Rowman & Littlefield Publishers.

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301--1309. Association for Computational Linguistics.

Carbonell, J. and Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335--336. ACM.

Carlsson, M. and Rooth, D.-O. (2007). Evidence of ethnic discrimination in the swedish labor market using experimental data. *Labour Economics*, 14(4):716--729.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. In *In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*.

Chakraborty, A., Messias, J., Benevenuto, F., Ghosh, S., Ganguly, N., and Gummadi, K. P. (2017). Who makes trends? understanding demographic biases in crowdsourced recommendations. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*.

Chen, H. and Karger, D. R. (2006). Less is more: probabilistic models for retrieving fewer relevant documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 429--436. ACM.

Chen, L., Wu, W., and He, L. (2016). *Personality and Recommendation Diversity*, pages 201--225. Springer International Publishing, Cham.

Chen, X., Wang, Y., Agichtein, E., and Wang, F. (2015). A comparative study of demographic attribute inference in twitter. In *In Proceedings of the 9th International AAAI Conference on Weblogs and Social Media*.

Cheng, A., Evans, M., and Singh, H. (2009). Inside twitter: An in-depth look inside the twitter world. *Report of Sysomos, June, Toronto, Canada*.

Comarela, G., Crovella, M., Almeida, V., and Benevenuto, F. (2012). Understanding factors that affect response rates in twitter. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*.

Cotter, D. A., Hermsen, J. M., Ovadia, S., and Vanneman, R. (2001). The glass ceiling effect. *Social forces*, 80(2):655--681.

Culotta, A., Kumar, N. R., and Cutler, J. (2015). Predicting the demographics of twitter users from website traffic data. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*.

Cunha, E., Magno, G., Almeida, V., Gonçalves, M. A., and Benevenuto, F. (2012). A gender based study of tagging behavior in twitter. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, HT '12, pages 323--324, New York, NY, USA. ACM.

Datta, A., Tschantz, M. C., and Datta, A. (2015). Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92--112.

De Choudhury, M., Sharma, S. S., Logar, T., Eekhout, W., and Nielsen, R. C. (2017). Gender and cross-cultural differences in social media disclosures of mental illness. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 353--369, New York, NY, USA. ACM.

Diakopoulos, N. and Koliska, M. (2016). Algorithmic transparency in the news media. *Digital Journalism*.

Dovidio, J. F. and Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological science*, 11(4):315--319.

Dutta, S. (2010). What's your personal social media strategy. *Harvard business review*, 88(11):127--130.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214--226. ACM.

Edelman, B., Luca, M., and Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, 9(2):1--22.

Fan, H., Cao, Z., Jiang, Y., Yin, Q., and Doudou, C. (2014). Learning deep face representation. *arXiv preprint arXiv:1403.2802*.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 259--268, New York, NY, USA. ACM.

Felton, E. and Lyons, T. (2016). The administration's report on the future of artificial intelligence. http://whitehouse.gov/blog/2016/10/12/administrations-report-future-artificial-intelligence.

Freitas, C., Benevenuto, F., Ghosh, S., and Veloso, A. (2015). Reverse engineering socialbot infiltration strategies in twitter. In *Proceedings of the 2015 IEEEACM International Conference on Advances in Social Networks Analysis and Mining*.

Ge, Y., Knittel, C. R., MacKenzie, D., and Zoepf, S. (2016). Racial and gender discrimination in transportation network companies. Technical report, National Bureau of Economic Research.

Gilbert, E., Bakhshi, S., Chang, S., and Terveen, L. (2013). "i need to try this"?: A statistical overview of pinterest. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM.

Goel, S., Rao, J. M., Shroff, R., et al. (2016). Precinct or prejudice? understanding racial disparities in new york city's stop-and-frisk policy. *The Annals of Applied Statistics*, 10(1):365--394.

Goh, G., Cotter, A., Gupta, M., and Friedlander, M. P. (2016). Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, pages 2415--2423.

Gollapudi, S. and Sharma, A. (2009). An axiomatic approach for result diversification. In *ACM WWW*.

Graells-Garrido, E., Lalmas, M., and Menczer, F. (2015). First women, second sex: gender bias in wikipedia. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*.

Guynn, J. (2015). Google photos labeled black people 'gorillas'. *USA Today, July*.

Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., and Wilson, C. (2017). Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 1914--1933, New York, NY, USA. ACM.

Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1--33.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. *Machine Learning and Knowledge Discovery in Databases*, pages 35--50.

Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., and Strohmaier, M. (2016). Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the 25th International Conference on World Wide Web*.

Koumchatzky, N. and Andryeyev, A. (2017). Using deep learning at scale in twitter's timelines. http://blog.twitter.com/engineering/en_us/topics/insights/2017/using-deep-learning-at-scale-in-witters-timelines.html.

Kricheli-Katz, T. and Regev, T. (2016). How many cents on the dollar? women and men in product markets. *Science advances*, 2(2):e1500599.

Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., and Karahalios, K. (2017). Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing*.

Kulshrestha, J., Kooti, F., Nikravesh, A., and Gummadi, P. K. (2012). Geographic dissection of the twitter network. In *In Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*.

Liu, W. and Ruths, D. (2013). What's in a name? using first names as features for gender inference in twitter. In *AAAI Spring Symposium Series*, volume 13, page 01.

Luong, B. T., Ruggieri, S., and Turini, F. (2011). k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 502--510, New York, NY, USA. ACM.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415--444.

Merluzzi, J. and Dobrev, S. D. (2015). Unequal on top: Gender profiling and the income gap among high earner male and female professionals. *Social science research*, 53:45--58.

Messias, J., Benevenuto, F., Weber, I., and Zagheni, E. (2016). From migration corridors to clusters: The value of google+ data for migration studies. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

Messias, J., Schmidt, L., Rabelo, R., and Benevenuto, F. (2013). You followed my bot! transforming robots into influential users in twitter. *First Monday*, 18(7).

Messias, J., Vikatos, P., and Benevenuto, F. (2017). White, man, and highly followed: Gender and race inequalities in twitter. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, WI '17, New York, NY, USA. ACM.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J. N. (2011). Understanding the demographics of twitter users. In *In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*.

Morales, A., Borondo, J., Losada, J. C., and Benito, R. M. (2014). Efficiency of human activity on information spreading on twitter. *Social Networks*, 39:1--11.

Morstatter, F., Pfeffer, J., and Liu, H. (2014). When is it biased?: assessing the representativeness of twitter's streaming api. In *Proceedings of the 23rd International Conference on World Wide Web*.

Nilizadeh, S., Groggel, A., Lista, P., Das, S., Ahn, Y.-Y., Kapadia, A., and Rojas, F. (2016). Twitter's glass ceiling: The effect of perceived gender on online visibility. In *In Proceedings of the 10th International AAAI Conference on Weblogs and Social Media*.

Nunez, M. (2016). Former Facebook Workers: We Routinely Suppressed Conservative News. gizmodo.com/former-facebook-workers-we-routinely-suppressed-conser-1775461006.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*.

Ohlheiser, A. (2016). Three days after removing human editors, facebook is already trending fake news. washingtonpost.com/news/ the-intersect/wp/2016/08/29/a-fake-headline-about-megyn-kelly-was-trending-on-facebook/.

Oliver, M. L. and Shapiro, T. M. (2006). *Black wealth, white wealth: A new perspective on racial inequality.* Taylor & Francis.

Pedreshi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560--568. ACM.

Pennacchiotti, M. and Popescu, A.-M. (2011). A machine learning approach to twitter user classification. In *Proceedings of the 5th International AAAI Conference on Web and Social Media.*

Reis, J. C. S., Kwak, H., An, J., Messias, J., and Benevenuto, F. (2017). Demographics of news sharing in the u.s. twittersphere. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media.* ACM.

Ridgeway, G. and MacDonald, J. M. (2009). Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. *Journal of the American Statistical Association*, 104(486):661--668.

Romei, A. and Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(05):582--638.

Romero, D. M., Galuba, W., Asur, S., and Huberman, B. A. (2011). Influence and passivity in social media. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 18--33. Springer.

Ronson, J. (2016). *So you've been publicly shamed.* Riverhead Books (Hardcover).

Sharma, N. K., Ghosh, S., Benevenuto, F., Ganguly, N., and Gummadi, K. (2012). Inferring who-is-who in the twitter social network. *ACM SIGCOMM Computer Communication Review*, 42(4):533--538.

Sheth, S. K., Bell, J. S., Arora, N., and Kaiser, G. E. (2011). Towards diversity in recommendations using social networks.

Sloan, L., Morgan, J., Burnap, P., and Williams, M. (2015). Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PloS one*, 10(3):e0115545.

Sugimoto, C. R., Lariviere, V., Ni, C., Gingras, Y., Cronin, B., et al. (2013). Global gender disparities in science. *Nature*, 504(7479):211--213.

Sugimoto, C. R., Ni, C., and Larivière, V. (2015). On the relationship between gender disparities in scholarly communication and country-level development indicators. *Science and Public Policy*, 42(6):789--810.

Sweeney, L. (2013). Discrimination in online ad delivery. *Queue*, 11(3):10:10--10:29. ISSN 1542-7730.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24--54.

Todisco, M. (2014). Share and share alike: Considering racial discrimination in the nascent room-sharing economy. *Stan. L. Rev. Online*, 67:121--129.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the 4th International AAAI Conference on Web and Social Media*.

Victor, D. (2016). Microsoft created a twitter bot to learn from users. it quickly became a racist jerk. *The New York Times*.

Vikatos, P., Messias, J., Miranda, M., and Benevenuto, F. (2017). Linguistic diversities of demographic groups in twitter. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. ACM.

Wachs, J., Hannak, A., Voros, A., and Daroczy, B. (2017). Why do men get more attention? exploring factors behind success in an online design community. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM'17, Montreal, Canada.

Wagner, C., Graells-Garrido, E., Garcia, D., and Menczer, F. (2016). Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science*, 5(1):1.

Willis, J. and Todorov, A. (2006). First impressions making up your mind after a 100-ms exposure to a face. *Psychological science*, 17(7):592--598.

Wu, S., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Who says what to whom on twitter. In *Proceedings of the 20th International Conference on World Wide Web*.

Yin, Q., Cao, Z., Jiang, Y., and Fan, H. (2015). Learning deep face representation. US Patent 20,150,347,820.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017a). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1171--1180, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. (2017b). Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, AISTATS '17.

Zagheni, E., Garimella, V. R. K., Weber, I., and State, B. (2014). Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd International Conference on World Wide Web*.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, ICML '13, pages 325--333.

Ziegler, C.-N., McNee, S. M., Konstan, J. A., and Lausen, G. (2005). Improving recommendation lists through topic diversification. In *ACM WWW*.