Linguistic Diversities of Demographic Groups in Twitter

Pantelis Vikatos
University of Patras
Rio, Greece
vikatos@ceid.upatras.gr

Johnnatan Messias
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brazil
johnnatan@dcc.ufmg.br

Manoel Miranda
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brazil
manoelrmj@dcc.ufmg.br

Fabrício Benevenuto
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brazil
fabricio@dcc.ufmg.br

ABSTRACT

The massive popularity of online social media provides a unique opportunity for researchers to study the linguistic characteristics and patterns of user's interactions. In this paper, we provide an in-depth characterization of language usage across demographic groups in Twitter. In particular, we extract the gender and race of Twitter users located in the U.S. using advanced image processing algorithms from Face++. Then, we investigate how demographic groups (i.e. male/female, Asian/Black/White) differ in terms of linguistic styles and also their interests. We extract linguistic features from 6 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 differences in particular writing set of attributes. In addition, 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 unveils clear differences in the writing styles (and the topics of interest) of different demographic groups, with variation seen across both gender and race lines. We hope our effort can stimulate the development of new studies related to demographic information in the online space.

KEYWORDS

demographics, linguistics, Twitter analysis

ACM Reference format:
Pantelis Vikatos, Johnnatan Messias, Manoel Miranda, and Fabrício Benevenuto. 2017. Linguistic Diversities of Demographic Groups in Twitter. In *Proceedings of HT '17, Prague, Czech Republic, July 4–7, 2017,* 10 pages.
DOI: http://dx.doi.org/10.1145/3078714.3078742

1 INTRODUCTION

The number of users in online social networking sites, such as Facebook and Twitter increases each day. As of the third quarter of 2016, Facebook and Twitter have 1.79 billion[1] and 317 million[2] monthly active users, respectively, sharing content about their daily lives and things that happen around them. This massive popularity of online social media 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 group, demographics constitute a significant factor to cluster people and understand their behavior. Twitter provides a plethora of different information, e.g. posts, social connections. However, it lacks data about demographics such as gender, race, or age. We deal with this absence of this information using profile image as an input of deep learning algorithms on image processing. We are interested in extracting demographic status in a large scale and correlate it with available information on the social media. Twitter is a microblogging platform so the main way of communication and action is by posting texts (tweets). The use of natural language processing in these type of data can extract many features describing cognitive and user' personal concerns.

Many studies have used text analysis to study the user behavior in the online space [2, 6, 7, 11]. Our work provides a complementary perspective to these efforts, by providing a characterization of language usage (i.e. common phrases and topics of interest), but grouping users according to their gender and race. Our effort is motivated by previous studies that uses computational linguistics in order to extract patterns about demographic information [14], but our effort further explores race as a new demographic dimension. Our findings reveal significant differences between the linguistic content shared by female and male users as well as Asian, Black, and White and can be used for automatically categorization of Twitter users through their texts.

The main challenge is that users in Twitter are prone not to provide information about demographics. In our work, we crawled a large scale sample of active Twitter users and then we identify the gender and race of about 1.6 million users located in U.S by using Face++[3] [15, 30], 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

HT '17, Prague, Czech Republic

978-1-4503-4708-2/17/07…$15.00
DOI: http://dx.doi.org/10.1145/3078714.3078742

[1] https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/
[2] https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/
[3] http://www.faceplusplus.com

and image processing, can provide with high accuracy the gender, race, and even the age of an individual via his/her image. From the demographic recognized users, we gathered tweets of 304, 477 users to characterize linguistic patterns. Particularly, 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 concludes that 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.

The rest of the paper is organized as follow. Section 2 provides a review of the relevant literature. Then, Section 3 presents the Twitter and demographic dataset. After that, the analysis and discussion of linguistic differences and topic of interests are presented. Finally, the last section summarizes our results and offers some concluding remarks.

2 RELATED WORK

In this section, we review the related literature along two axes. First, we discuss the methodology used by efforts that measure demographic factors in Twitter. Then, we refer to studies that combine linguistic with demographic status.

2.1 Demographics in Social Media

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 [23]. After that, several efforts have arisen that investigate demographic information, in various social media, using different strategies for distinct purposes [4, 5, 18, 19, 29]. Particularly, in terms of text analysis, Cunha *et al.* [13] used Twitter data to analyze the difference between males and females in terms of generation of hashtags. Their results emphasize gender as factor able to influence the user's choice of specific hashtags to a specific topic.

Recent studies focused on demographics [4, 19, 21, 24, 25] 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.* [9] focus on demographic inference using namely profile self-descriptions and profile images. They categorize demographic status using as signals users' names, self-descriptions, tweets, social networks, and profile images to infer attributes as ethnicity, gender, and age. An alternative approach, Culotta *et al.* [12] declare that the demographic profiles of visitors to a website are correlated with the demographic profiles of followers of that website on the social network and propose a regression model to predict demographic attributes such as gender, age, ethnicity, education, income, and child status. More recently, An *et al.* [1] provide an accurate scheme in order to predict gender and race using the correlation of hashtags that are used in different demographic groups.

Finally, our effort uses the similar strategy to gather demographic information as Chakraborty *et al.* [8], but we investigate very different research questions as we focus on the linguistic analysis of demographic groups.

2.2 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.* [5] produce n-grams 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% acc) for gender classification. Also, Chen *et al.* [9] introduce a similar methodology for predicting gender, ethnicity, and age. However, using n-grams from the social neighbors, including followers and friends, and the distribution of 100 generated topics of LDA algorithm as the input of SVM classifier. Their results present that the performance of classification is much lower in terms of ethnicity and age. Gilbert *et al.* [17] present an interesting statistical overview in Twitter and Pinterest using textual analysis and comparing what users text on Pinterest to what they text on Twitter.

We mainly motivate our research based on Choudhury *et al.* [14] study which discover gender and cultural differences in Twitter. They correlate several linguistic features to mental illness. Our findings reinforce their observations about linguistic and topical differences against male and female users in Twitter and also contribute with a new analysis of race.

3 DEMOGRAPHIC INFORMATION DATASET

This section focuses on the procedure of data collection in order to extract useful inference about the discrimination of demographic status of a Twitter user. Our ultimate goal consists of gather demographic characteristics as gender and race as well as attributes about social behavior and tweet activity of active U.S. Twitter users. Next, we describe our steps to create this dataset and also discuss its main limitations.

3.1 Twitter dataset gathered

Our procedure uses the provided information from Twitter Stream API [4] in order to identify active Twitter users. We use a time window of three complete months from July to September 2016, collecting 341, 457, 982 tweets posted by 50, 270, 310 users.

Due to the fact that geographic coordinates are available on Twitter only for a limited number of users (i.e. $< 2\%$) [7], our strategy to identify U.S. Twitter users is based on the time zone information to retrieve users which are actually from the US as the methodology in previous efforts [8, 20] presented.

We filtered 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.

3.2 Crawling Demographic Information

The field of demographic status is not mandatory when a user registers in Twitter and, thus, the direct retrieval of gender, race, or even age is not feasible. There are several studies related to demographic information in Twitter that attempt to infer the user's gender from the user name [4, 19, 21, 23]. Also, some works use

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

pattern based methodology to identify age [27] in Twitter profile description using regular expressions *'25 yr old'* or *'born in 1990'*.

Here, we use a different strategy that allows us to extract the demographic dimension using the profile picture of each user. To do that, we needed to gather 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.

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 [15, 30] able to identify the gender (i.e. male and female) and race (limited to Asian, Black, and White) from recognized faces in images.

Table 1: 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	2 million
U.S. users with one face (Baseline)	1.6 million
U.S. users with crawled tweets	304 thousand

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 final dataset contains 1, 670, 863 users located in U.S. with identified demographic information. The phases of our data crawling and the amount of data discarded on each step are summarized in Table 1.

3.3 Baseline Dataset

In this section, we use the null model as our approach to estimate the statistical significance of the observed trend in given data. We compare the distribution of random samples created by the null model with the one of the original dataset and we measure the statistical significance.

Table 2 shows the distribution of gender and race in the dataset of the ≈ 1.6 million Twitter users between July and September 2016. To construct a null model, we create k random samples from the entire dataset (our crawled dataset containing 1.6 million users with demographic attributes), where each sample has exactly 304, 477 users. We choose this value for each sample size as it corresponds to the number of users we were able to gather tweets. For each sample, we count how many Whites are included. Then, the Z_{White} is computed as following:

$$Z_{White} = \frac{|U_{White}| - mean(|S_{White}|)}{std(|S_{White}|)} \tag{1}$$

where $mean(\cdot)$ is the mean and $std(\cdot)$ is the standard deviation of the values from multiple samples. We use the same equation for the other gender and race attributes. Table 3 presents the demographic distribution of 304, 477 users with linguistic attributes. The numbers in the parenthesis correspond the Z-values.

Table 2: Demographic distribution of 1.6 million users, our Baseline dataset.

Race (%)	Gender (%)		Total
	Male	Female	
Asian	7.24	10.61	17.85
Black	7.84	6.45	14.29
White	32.23	35.63	67.86
Total	47.31	52.69	100.00

Table 3: Demographic distribution of 304, 477 users with linguistic attributes. The numbers in the parenthesis correspond 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 4: Basic statistical descriptions of number of tweets with confidence intervals of 95% confidence level.

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

Intuitively, when the absolute value of Z-value becomes bigger (either positive or negative), the trend (more number or less number, respectively) is less likely observed by chance. In this work, we use 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 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 if we pick users randomly in our dataset, we would expect demographic groups with these proportions. Table 4 shows the statistical descriptions of number of tweets with 95% confidence level for each demographic group.

3.5 Extraction of Topics

We extracted the information about topics of interests for active users using the *Who Likes What*[5] web service [3]. The produced topics are derived from the list of the friends (other users the user is following) of each user. Then, we sort the produced topics based on their frequency to conclude the 20 most common topics from the Twitter users, including them as Binary variables. We manually cleaned several top topic labels following the same procedure as [24]. Therefore, we merged 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 5 presents a list of the 20-top topics and the merged sub-topics in each one as well as the number of users that belong to them.

3.6 Linguistic Measures

To quantify gender and race dimensions in the language of Twitter users, we use the 2015 version of the psycholinguistic lexicon Linguistic Inquiry and Word Count (LIWC) [28]. Since LIWC has been proposed, it has been widely used for a number of different tasks, including sentiment analysis [26] and discourse characterization in social media platforms [11]. The features are categorized into 3 main categories, (1) affective attributes, (2) cognitive attributes, and (3) linguistic style attribute as Choudhury *et al.* [14] propose. For this work, we considered 36 features from LIWC categorized into 6 groups in order to find the main differences across each demographic group.

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 focuses in present 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 Data 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. 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 ± 0.015% for gender and 85.97±0.024% for race, with a confidence interval of 95%. Recent efforts have used Face++ for similar tasks and reported fairly well confidence in manual inspections [1, 8, 31]. Our dataset may contain fake accounts and bots as previous studies provide evidence for a non-negligible rate of fake accounts [16, 22] in Twitter.

Finally, we note that our approach to identify users located in U.S. may bring together some users located in the same 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 [10].

4 LINGUISTIC DIFFERENCES

In order to show how demographic groups differ from each other in both gender and race domains, this section presents the difference between demographic groups across various linguistic categories. Table 6 shows the linguistic features extracted from LIWC into 6 categories (affective attributes, cognitive attributes, lexical density and awareness, temporal references, social and personal concerns, and interpersonal focus).

Figure 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 instance is calculated as the average of the absolute difference of each feature that comprises this group. This shows in which linguistic categories the analyzed users differ the most. The amount of users considered in each group were the same.

Figure 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. In counterpart, from the race domain, the differences tend to be higher in affective attributes.

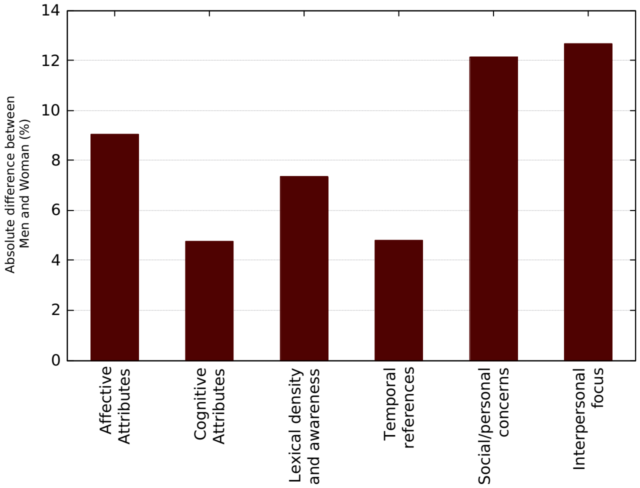


Figure 1: Mean absolute differences between male and female users per the various categories of linguistic measures

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 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

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

Table 5: 20-top Topics of user’s interests

Topic	Sub-Topics	Total
Celebrities	celebrities, famous, stars, celebs, celebrity, star, celeb	1, 319, 765
Artists	musicians, singers, artist, singer, musician, rappers, bands	731, 370
World	world, earth, hollywood, usa, canada, texas, international, nyc, country, city, boston, san francisco, france, america, los angeles, brasil, london, india	654, 555
Music	music, pop, hip hop, rap, gospel, hiphop	463, 451
Fun	fun, funny, humor, lol, laugh	415, 113
Entertainment	entertainment	371, 503
TV	tv, television	369, 440
Info	info, information	297, 705
Sports	sports, football, basketball, baseball, soccer, futbol, basket, martial arts, sport, mma, golf, cricket, boxing, motorsports, f1, racing	296, 652
Media	sports news, tech news, newspapers, music news, breaking news, world news, news media, radio, internet, social media, youtube, sports media, magazines, magazine	293, 206
Life	life, lifestyle, health, healthcare, fitness, food, style, smile, drink	278, 348
Actors	actors, actresses, actress, actor	267, 626
Bloggers	bloggers, blogs, blog	230, 347
Technology	technology, tech, iphone, digital, geek, software, computer, electronic, android, xbox, mac, gadgets, programming, geeks	208, 739
Movie	movie, movies, film, films	203, 577
Writers	writers	189526
Organizations	organizations, nfl, nba, mlb, nhl, ufc, lfc, lgbt	178, 030
Business	business, biz, businesses	171, 759
Politics	politics, government, political, politicians, politician	110, 367
Companies	companies, apple, company, microsoft, google	79, 528

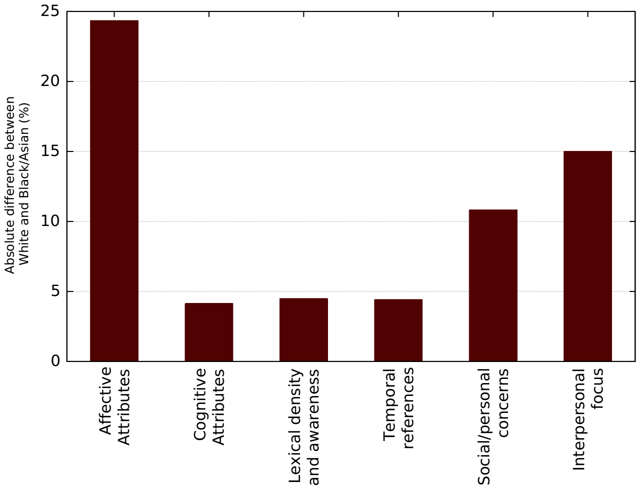


Figure 2: Mean absolute differences between White and Black/Asian users combined per the various categories of linguistic measures

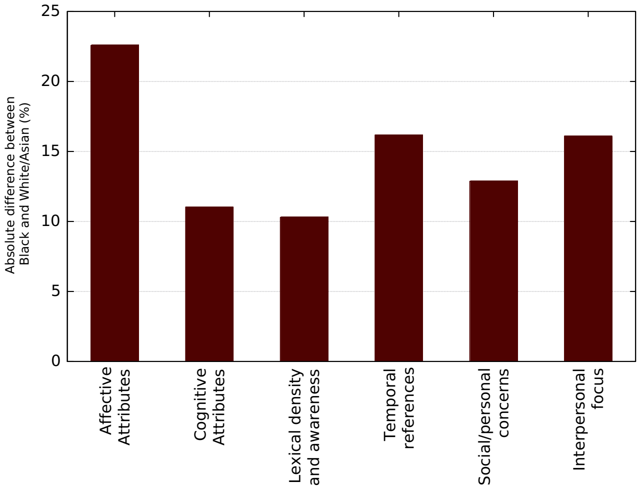


Figure 3: Mean absolute differences between Black and White/Asian users combined per the various categories of linguistic measures

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 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 based on Wilcoxon rank sum significance tests. p-values are

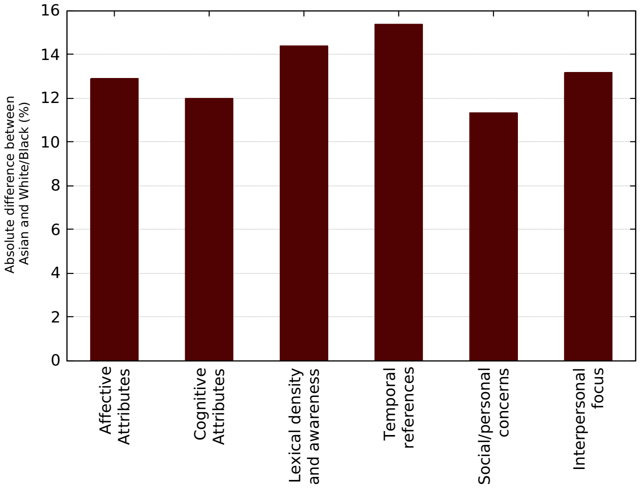


Figure 4: Mean absolute differences between Asian and White/Black users combined per the various categories of linguistic measures

represented in asterisks scale using * as significant ($0.1 < p \leq 0.5$), ** very significant($0.001 < p \leq 0.01$), ***($0.0001 < p \leq 0.001$) and ****($p < 0.001$) extremely significant. As Table 6 presents, females tend to use anxiety ($z = -74.534$) and sadness ($z = -74.394$) terms and phrases. On the other hand, males express with 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 ($z = -15.118$)

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 notorious values shown in Table 6, 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 ($z = 62.137$), 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.

When talking about Social/Personal concerns, there is a higher presence of Black people in the features from this class, noticeably in family ($z = 86.721$), social ($z = 90.830$), religion ($z = 85.163$), and body ($z = 86.903$).

The Interpersonal Focus feature set reveal 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 8 & Table 9 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 8 phrases expressing negation are in the top 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 9, the trend of using negation phrases also repeat 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. Table 8 and Table 9 show differences regarding the way of writing of each demographic group and reveal interesting characteristics about the difference from one to another.

5 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 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

Table 6: Differences between tweets from male and female users based on linguistic measures. $\mu(male)$ and $\mu(female)$ are the median values of feature for male and female, respectively. Statistical significance is count based on Wilcoxon rank sum tests. p-values are represented in asterisks scale using * as significant $(0.1 < p \leq 0.5)$, ** very significant $(0.001 < p \leq 0.01)$, $^{***}(0.0001 < p \leq 0.001)$ and $^{****}(p < 0.001)$ extremely significant.

	$\mu(male)$	$\mu(female)$	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***

specifically, the gender difference between topics varies among males and females.

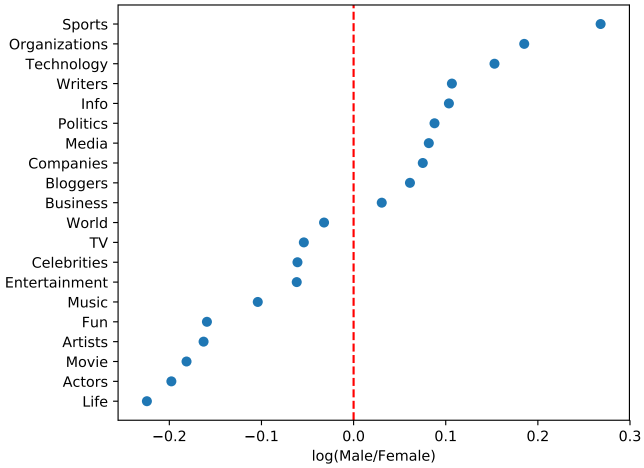


Figure 5: Gender interests: Blue dots represent the gender interests for the 20-top popular topics.

In a similar way, we present the race distribution for the 20-top topics of Asian, Black, and White users in Figure 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. Therefore, we have randomly selected 45, 398 users for each race to study their topic interests. Users from different races may also vary in interests and preferences. Figure 6-a shows that White users have more interest in politics, writers, and organizations than Asians. However, Asians prefer more artists, actors, and music topics than Whites. Figure 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, when we look at Figure 6-c, Asians interest more for movie, companies, and technology topics than Blacks. On other hand, Blacks prefer more business, sports, and organizations than Asians.

6 CONCLUSION

The results presented in this paper allow us to conclude that there are clear differences in the way of writing across different demographic groups in both gender and race domains. Our main contribution relies on characterizing the differences in the way of writing for each group pointing the most important linguistic aspects for a specific gender and race. Through the analysis of mean absolute differences amongst linguistic features between each demographic group, we were able to identify those which affective attributes were more present 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.

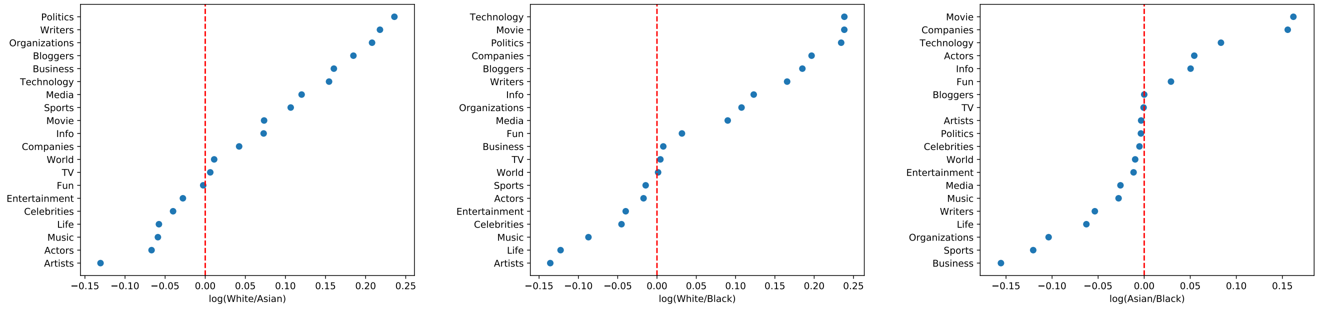
Another interesting conclusion is based on the most common phrases encountered on each group and their position ranking when

Table 7: 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 present 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

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 slangs, which is common in an environment like Twitter, appears in these frequent phrases too. When we compare the difference between the groups, we find interesting trends, like the higher interest in signs by females than by males.

By analyzing topic interests, we found that each demographic group tends to have its own preferences over the information they share. For instance, we found that males are more into sports,



(a) White vs Asian

(b) White vs Black

(c) Asian vs Black

Figure 6: Race interests: Blue 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.

Table 8: 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

organizations, and 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 higher interest for movie, companies, and technology when compared to Blacks.

There are some future directions we would like to pursue next. First, we plan 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. Also, our will is to extend this work correlating demographic aspects with the social behavior, e.g. number of followers, listed, etc. In addition, we plan to examine the speed of tweets that are propagated through a specific demographic group.

ACKNOWLEDGMENTS

This work was partially supported by the project FAPEMIG-PRONEX-MASWeb, Models, Algorithms and Systems for the Web, process number APQ-01400-1 and grants from CNPq, CAPES, Fapemig, and Humboldt Foundation.

REFERENCES

[1] Jisun An and Ingmar Weber. 2016. #greysanatomy vs. #yankees: Demographics and Hashtag Use on Twitte. In *Proceedings of the 10th International AAAI Conference on Web and Social Media*. 523–526.

[2] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. 2009. Characterizing user behavior in online social networks. In *IMC '09: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*. ACM, New York, NY, USA, 49–62. DOI: http://dx.doi.org/10.1145/1644893.1644900

[3] Parantapa Bhattacharya, Muhammad Bilal Zafar, Niloy Ganguly, Saptarshi Ghosh, and Krishna P. Gummadi. 2014. Inferring User Interests in the Twitter Social Network. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*. ACM, New York, NY, USA, 357–360.

[4] Cameron Blevins and Lincoln Mullen. 2015. Jane, John... Leslie? a historical method for algorithmic gender prediction. *Digital Humanities Quarterly* 9, 3 (2015).

[5] John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1301–1309.

[6] Meeyoung Cha, Fabrício Benevenuto, Hamed Haddadi, and Krishna P. Gummadi. 2012. The world of connections and information flow in Twitter. *IEEE Transactions on Systems, Man and Cybernetics - Part A* 42, 4 (2012), 991–998.

[7] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM'10)*. Washington DC, USA.

[8] Abhijnan Chakraborty, Johnnatan Messias, Fabricio Benevenuto, Saptarshi Ghosh, Niloy Ganguly, and Krishna P. Gummadi. 2017. Who Makes Trends? Understanding Demographic Biases in Crowdsourced Recommendations. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM'17)*. Montreal, Canada.

[9] Xin Chen, Yu Wang, Eugene Agichtein, and Fusheng Wang. 2015. A comparative study of demographic attribute inference in twitter. In *In Proceedings of the 9th International AAAI Conference on Weblogs and Social Media.*

[10] Alex Cheng and Mark Evans. 2009. Inside Twitter: An In-Depth Look Inside the Twitter World. (2009).

Table 9: 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	-	-

[11] Denzil Correa, Leandro Araújo Silva, Mainack Mondal, Fabrício Benevenuto, and Krishna P. Gummadi. 2015. The Many Shades of Anonymity: Characterizing Anonymous Social Media Content. In *In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM'10).*

[12] Aron Culotta, Nirmal Ravi Kumar, and Jennifer Cutler. 2015. Predicting the Demographics of Twitter Users from Website Traffic Data.. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence.* 72–78.

[13] Evandro Cunha, Gabriel Magno, Virgilio Almeida, Marcos André Gonçalves, and Fabricio Benevenuto. 2012. A Gender Based Study of Tagging Behavior in Twitter. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media (HT '12).* ACM, New York, NY, USA, 323–324. DOI: http://dx.doi.org/10.1145/2309996.2310055

[14] Munmun De Choudhury, Sanket S. Sharma, Tomaz Logar, Wouter Eekhout, and René Clausen Nielsen. 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).* ACM, New York, NY, USA, 353–369.

[15] Haoqiang Fan, Zhimin Cao, Yuning Jiang, Qi Yin, and Chinchilla Doudou. 2014. Learning deep face representation. *arXiv preprint arXiv:1403.2802* (2014).

[16] Carlos Freitas, Fabricio Benevenuto, Saptarshi Ghosh, and Adriano Veloso. 2015. Reverse Engineering Socialbot Infiltration Strategies in Twitter. In *Proceedings of the 2015 IEEEACM International Conference on Advances in Social Networks Analysis and Mining.*

[17] Eric Gilbert, Saeideh Bakhshi, Shuo Chang, and Loren Terveen. 2013. ”I Need to Try This”?: A Statistical Overview of Pinterest. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, USA, 2427–2436.

[18] Aniko Hannak, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr. In *20th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2017).* Portland, OR.

[19] Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. 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.* 53–54.

[20] Juhi Kulshrestha, Farshad Kooti, Ashkan Nikravesh, and P Krishna Gummadi. 2012. Geographic Dissection of the Twitter Network.. In *In Proceedings of the 6th International AAAI Conference on Weblogs and Social Media.*

[21] Wendy Liu and Derek Ruths. 2013. What's in a Name? Using First Names as Features for Gender Inference in Twitter.. In *AAAI Spring Symposium Series,* Vol. 13. 01.

[22] Johnnatan Messias, Lucas Schmidt, Ricardo Rabelo, and Fabrício Benevenuto. 2013. You followed my bot! Transforming robots into influential users in Twitter. *First Monday* 18, 7 (July 2013).

[23] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. Understanding the Demographics of Twitter Users.. In *In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media,* Vol. 11. 5th.

[24] Shirin Nilizadeh, Anne Groggel, Peter Lista, Srijita Das, Yong-Yeol Ahn, Apu Kapadia, and Fabio Rojas. 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.*

[25] Julio C. S. Reis, Haewoon Kwak, Jisun An, Johnnatan Messias, and Fabricio Benevenuto. 2017. Demographics of News Sharing in the U.S. Twittersphere. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media (HT '17).* ACM, New York, NY, USA.

[26] Filipe N. Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science* 5, 1 (2016), 1–29.

[27] Luke Sloan, Jeffrey Morgan, Pete Burnap, and Matthew Williams. 2015. Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PloS one* 10, 3 (2015), e0115545.

[28] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[29] Johannes Wachs, Aniko Hannak, Andras Voros, and Balint Daroczy. 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.

[30] Qi Yin, Zhimin Cao, Yuning Jiang, and Haoqiang Fan. 2015. Learning Deep Face Representation. (Dec. 3 2015). US Patent 20,150,347,820.

[31] Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, and Bogdan State. 2014. Inferring International and Internal Migration Patterns from Twitter Data. In *Proceedings of the 23rd International Conference on World Wide Web.*