Demographics of News Sharing in the U.S. Twittersphere

Julio C. S. Reis
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
julio.reis@dcc.ufmg.br

Haewoon Kwak
Qatar Computing Research Institute
Doha, Qatar
haewoon@acm.org

Jisun An
Qatar Computing Research Institute
Doha, Qatar
jan@hbku.edu.qa

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

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

ABSTRACT

The widespread adoption and dissemination of online news through social media systems have been revolutionizing many segments of our society and ultimately our daily lives. In these systems, users can play a central role as they share content to their friends. Despite that, little is known about news spreaders in social media. In this paper, we provide the first of its kind in-depth characterization of news spreaders in social media. In particular, we investigate their demographics, what kind of content they share, and the audience they reach. Among our main findings, we show that males and white users tend to be more active in terms of sharing news, biasing the news audience to the interests of these demographic groups. Our results also quantify differences in interests of news sharing across demographics, which has implications for personalized news digests.

KEYWORDS

Online News; Demographics; News Sharing; Social Media; Twitter

ACM Reference format:
Julio C. S. Reis, Haewoon Kwak, Jisun An, Johnnatan Messias, and Fabrício Benevenuto. 2017. Demographics of News Sharing in the U.S. Twittersphere. In *Proceedings of HT '17, Prague, Czech Republic, July 4–7, 2017,* 10 pages.
DOI: http://dx.doi.org/10.1145/3078714.3078734

1 INTRODUCTION

In recent years, with the huge success of Twitter and Facebook, social media has become one of the most important channels in news diffusion. In particular, Twitter's unique concepts of asymmetric "follow" and "retweet", which were later adopted by Facebook, allow users to follow each other's updates and propagate interesting pieces of information quickly and broadly [24]. Such great power to disseminate information embedded in social media naturally has attracted the news media. As a result, a majority of U.S. adults (62%) get news mostly on social media, according to a new survey by Pew Research Center [10].

Along with their traditional channels, news media manage their presence in social media by creating Twitter accounts and publishing tweets containing URLs that link their news media sites. For those accounts, it is clearly visible who the audience is – their followers. Furthermore, as any Twitter user can share URLs to news media web sites, Twitter users exposed to news media's tweets through retweets can also be visible and accounted as audience. We call these users *news spreaders* in the rest of this paper. This form of sharing of news URLs has long been a pervasive practice in social media, but its role and impact are relatively unexplored.

In this work, we characterize news spreaders in Twitter along three dimensions: 1) their demographics (who they are), 2) their news shared (what they share), and 3) their impact (why they are important). To this end, the inference of demographics of Twitter users is essential. Among various techniques that have been proposed [27], we use state-of-the-art techniques to locate Twitter users and infer their demographics based on profile photos.

Through a longitudinal data collection of news spreaders and their URL sharing behavior of five popular global news media, we test how similar news URL sharing is to typical URL sharing in terms of demographics of spreaders. We find a statistically significant trend that white males participate more in news URL sharing than other race-gender groups. This suggests that news spreaders have unique characteristics, which cannot be easily perceived for typical URL spreaders in Twitter. Thus, our work is essential to understand news spreaders correctly.

We then answer the above research questions. First, we examine demographics of news spreaders. By comparing the followers of news media accounts, we discover huge differences in terms of race-gender demographics. This suggests that we need to have a broader definition of the exposure of the news media on social media that are not only a set of followers [1] but also news spreaders. Second, we examine what kinds of news are shared by news spreaders. The properties of the pieces of news are defined along three dimensions: topics, author's (journalist's) gender and race, and linguistic analysis [33] of news headlines. These three dimensions have been discovered as important factors in news reading/sharing behavior [34, 38]. Finally, we answer how important news spreaders are for news media from the perspective of audience expansion: 1) about 59% of news spreaders do not follow news media accounts in Twitter; 2) the audience brought by the spreaders is much bigger than that of the original followers of the news media; 3) in addition

HT '17, Prague, Czech Republic

978-1-4503-4708-2/17/07…$15.00
DOI: http://dx.doi.org/10.1145/3078714.3078734

to that the demographics of the spreaders and those of the followers are quite different, the followers of the spreaders are also substantially different from the followers of the news sources in terms of demographics. In other words, the spreaders play an important role in expanding the audience of news in Twitter, which would otherwise be very limited. Lastly, we find that the demographics of news spreaders are related to the popularity of news.

Our contributions are three-fold: 1) by using a combination of state-of-the-art techniques, we investigate in details aspects of the audience of news media in Twitter, which has been considered as in-house data so far; 2) we suggest a robust statistical framework to test the news URL sharing behavior by comparing it with typical URL sharing behavior; and 3) Our findings show that news media should understand spreaders and their followers to capture the complete picture of their presence in news media. News media's direct followers are only the tip of the iceberg of their audience in Twitter in terms of volume and demographics.

The rest of the paper is organized as follows. Section 2 briefly surveys related efforts. Then, we present our experiment methodology and the data gathered. The next three sections cover our results. We conclude the paper by discussing implications from our findings as well as presenting directions for future work.

2 RELATED WORK

In this Section, we review existing work related to news sharing along two main dimensions.

2.1 News Sharing and Propagation

Social media services have made personal contacts and relationships more visible and quantifiable than ever before. Users interact by following each others' updates and passing along interesting pieces of information to their friends. This kind of word-of-mouth propagation occurs whenever a user forwards a piece of information to her friends, making users a key element in this process. Not surprisingly, a number of efforts have attempted to quantify and characterize information spread in social networks as well as the role users play in such propagation [11, 12, 36, 37, 42]. For example, Rodrigues et al. [36] showed that retweets are responsible for increasing the audience of URLs by about 2 orders of magnitude. As social media became an important channel in news diffusion, some recent research efforts attempted to investigate how news are shared in these systems. Next, we detail a few approaches that provides news sharing and propagation.

Naveed et al. [29] showed that bad news tends to spread faster in systems like Twitter. In this same year, also with the use of this same social media, Armstrong et al. [5] analyzed how online media companies employ men and women in Twitter feeds and how it connects to portrayals in news. In particular, the authors looked at how mentions of men and women on Twitter may influence mentions in news stories (e.g. newspaper, television). Through the content analysis of newspaper and television tweets at different granularity (i.e. local, regional and national), they found that male mentions were more likely to appear in national news than in regional or local news and more often than female mentions in the print media than on television.

A recent effort [7] has tackled the question "Why are some news articles shared more than others?". They showed that story importance cues are relevant in driving social sharing and that certain topics (i.e. stories about politics, accidents, disasters, and crime) were less shared. Some topics can be shared in order to improve the users' reputation. This dynamic media attention has inspired other recent studies [3]. Bright et al. [7], compare different social networks platforms and showed that some kind of news are shared more in one network than the others (e.g. economy news on LinkedIn).

Unlike previous works, our effort focuses on understanding the dynamics of news sharing on Twitter of each demographic group. Thus, to the best of our knowledge, this is the first effort that investigates intersection between news sharing and demographic information of users, including how these aspects are related.

2.2 Demographics in Social Media

Mislove et al. [27] was one of the first researchers that analyze demographic characteristics of Twitter users considering a geographical perspective (i.e. how the demographics vary across different U.S states). After that, several efforts have arisen that investigate demographic information, in various social media, using different strategies for distinct purposes [6, 8, 22]. Particularly, researchers are jointly applying computer science and statistical techniques to support sociological studies using large-scale social media datasets. These studies can range from a simple characterization of to the investigation of more complex causes, including to raising attention to the different levels at which gender biases can manifest themselves on the web [41].

In [16] the authors used Twitter data to analyze the difference between men and women behavior in terms of dynamics in free tagging environments. The results obtained present gender distinctions in the use of Twitter hashtags, emphasizing it as a social factor influencing the user's choice of specific hashtags on a specific topic. Still about tags (or hashtags), recently, the work presented in [4] explored their use by different demographic groups. The demographic characteristics of each user were obtained using *Face++* and the Twitter user's profile picture. The results showed that, although there are more popular hashtags that are commonly used, there are also many group-specific hashtags with non-negligible popularity. Besides that, the researchers show that the strategy of getting demographic data from *Face++* is reliable and provides accurate demographic information for gender and race, encouraging the application of this strategy in other recent efforts [13]. We use a similar strategy to gather demographic information.

Nilizadeh et al. [31] explore gender inequalities in Twitter, showing that gender may allow inequality to persist in terms of online visibility. Looking at Pinterest, Gilbert et al. [20] investigated what role gender plays in the website's social connections. The results highlight a major difference between female and male users regarding their motivations for using this social media. They found that being female means more repins (i.e., more shared content), but fewer followers in comparison with Twitter. Gender differences has also been explored in terms of social media disclosures of mental illness [17].

News Media	#Shares	#Authors	Screen name	#Followers
New York Times	14,505	1,165	@nytimes	1,141
Reuters	4,712	485	@Reuters	1,259
The Guardian	4,457	844	@guardian	1,620
Wall Street Journal	1,379	313	@WSJ	1,445
BBC News	1,144	190	@BBCBreaking	1,130

Table 1: Data collection by news source.

Race (%)	Gender (%)		**Total:**
	Male	Female	
Asian	5.29	6.05	11.34
AF-AM	6.09	3.80	9.89
White	43.46	35.31	78.77
Total:	54.84	45.16	100.00

Table 2: Demographic distribution of news spreaders.

More recently, An et al. [2] examined the news consumption in South Korea (from Daum News portal). The authors analyzed on a large scale the differences in news consumption from a demographic perspective. Through a multidimensional analysis of gender and age differences in news consumption, they quantify such differences along four distinct dimensions: actual news items, topic, issue, and angle. The top 30 news items for each gender and age group in Daum News were used and the demographics information were obtained through the website itself. Overall, focus mainly on quantifying and explaining differences in news consumption.

More broadly, most of the previous efforts attempt to quantify differences in gender behavior and inequalities in different social media or news systems. Our effort is the first of its kind to provide a characterization of news sharing across different demographic groups. Thus our effort is complementarity to the existing ones.

3 METHODOLOGY

In order to understand demographics of news sharing in Twitter, first we define our strategy for data collection. Then, we define our strategies for inference of demographic information of each individual Twitter user and collection of information such as category and authors of the news, and followers of each of the news media on Twitter. Our ultimate goal, in this section, consists of reporting our baseline for comparison in order to verify the statistical significance of the results. Next, we briefly describe the methodology adopted for this work, including a discussion of its main limitations.

3.1 Gathering Twitter

For this work, we gathered the 1% random sample of all tweets, through the Twitter Streaming API [1], along a 3 months period, from July to September, 2016. Specifically, we considered only tweets (and retweets) that contain at least one URL and have been shared by U.S. users. We understand that users who share URLs may present a slight difference in behavior compared to others, so, considering our research objective, we only select this set of American users. Besides that, as we are interested in analyzing demographic characteristics, it is important to study users from the same place. For this reason, we consider only U.S. users, filtered by timezone. In total, we retrieved 11,790,679 tweets posted by 11,770,273 U.S. users. From this initial dataset, we infer demographics information about users and build: (i) our news sharing dataset, used in the execution of our experiments, and (ii) our baseline dataset.

3.2 Inferring Demographics Information

In the literature, several studies present strategies for inference of gender, race, and age. Some efforts attempt to infer the gender of a user from her name [22, 25, 27], or the age from Twitter profile descriptions [39], by using patterns like '*like 25 yr old*' or '*born in 1990*'. However, in some cases the number of unrealized inferences (e.g. for lack of information) is high (Liu and Ruths [25] reported 66% users in their dataset did not have a proper name).

To overcome such limitation, in this work, we use the profile picture's URLs of all users in our dataset and use the *Face++* API [2], a face recognition platform based on deep learning [43], to infer the gender (i.e., male or female), race (limited to Asian, Black[3], and White) and age information from the recognized faces in the profile images. We discarded users whose profile pictures do not have a recognizable face or have more than one face, according to *Face++*. Our final dataset contains 937,308 unique users located in U.S. with identified demographic information, which are gender, race, and age by *Face++*.

3.3 Shared News Dataset

To focus on news sharing in Twitter, we filtered only tweets that shared news URLs from important and different news sources (i.e. BBC News[4], The New York Times[5], Reuters Online[6], The Wall Street Journal[7], The Guardian[8] and BBC News[9]), known worldwide. All these news sites appear among the most popular ones in the world, according to Alexa.com.[10] Simultaneously, we gathered information from users who posted each of the tweets including demographic information from *Face++*, as detailed above. From news URLs, we crawled information about them including, title, text, principal image (link - when there is one), authors (when there is one) and date. Lastly, Table 1 shows the dataset used in this work containing 26,211 unique news articles shared by 16,382 unique users. We note that The New York Times is the most widely shared news media in Twitter, in comparison with those news sites considered. Table 2 shows the demographic decomposition of those 16,382 users who shared news URLs.

3.4 Inferring News Category

In order to infer the categories of the news articles, we use meta information embedded in the news URLs. News media usually have several news sections, such as Politics, Sports, or World News, and group their news articles by these sections. By looking at which section a news article belongs to, we can infer a topical category

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

[2] https://www.faceplusplus.com

[3] We called *African-American (AF-AM)* in the rest of this paper.

[4] http://www.bbc.com

[5] http://www.nytimes.com

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

[7] http://www.wsj.com/

[8] https://www.theguardian.com

[9] http://www.bbc.com

[10] http://www.alexa.com/topsites/category/News

of the news articles. The section information is often embedded in news URL. For example, the URL http://www.nytimes.com/2016/07/02/us/politics/loretta-lynch-hillary-clinton-email-server.html represents that the news article is about "Politics". We parsed all News URLs and extracted the topic information. The New York Times, The Guardian, and BBC adopt the above mentioned strategy for their URLs, and thus we simply parse their URL and infer the topic of a given news article. Reuters and The Wall Street Journal do not have category information in their URLs, however, the news articles have the category information. Thus, we collected news articles and extracted category information by parsing HTML files. We successfully inferred the topical categories of 93.3% (24,466) of news articles. Figure 1 shows the proportion of the top 10 most significant news categories. We find that "World" is the most "shared" category (21.16%), similar to the results in [34].

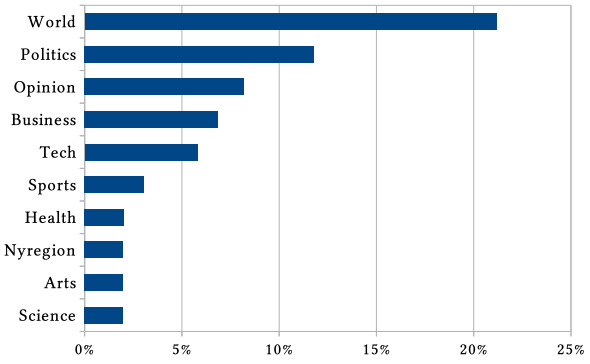

Figure 1: Top 10 most significant news categories.

3.5 Finding Journalists in Twitter

We aim to collect demographics of the authors of news articles in our dataset. Figure 2 shows the procedure for creating an author dataset. For each news URL, we collect its title, text, principal image, authors, and date by parsing the original web page. Then, we search and collect the Twitter profiles of the authors if they have Twitter accounts. Then, we infer those authors' demographic characteristics using *Face++* (see Section 3.2). Table 1 shows the number of authors for each news media. As expected, the largest number of names of distinguished authors we have gathered are from the The New York Times news media, which had the largest number of news shared in Twitter in our dataset.

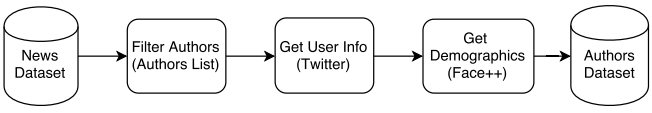

Figure 2: Strategy for collecting news authors.

3.6 Collecting Followers of News Media in Twitter

For each news source, we collected their followers in Twitter. Again, we infer their demographics by *Face++*. Table 1 presents the total

Race (%)	Gender (%)		**Total:**
	Male	Female	
Asian	7.07	10.33	17.40
AF-AM	8.52	6.93	15.45
White	31.97	35.18	67.15
Total:	47.56	52.44	100.00

Table 3: Demographic distribution of users in the Baseline dataset.

of gathered news media followers in Twitter, including the screen name used for collection. On average, we retrieved 1,319 followers by news source.

3.7 Baseline Dataset

A null model is widely used to estimate the statistical significance of the observed trend in given data. As the null model is randomly generated data that preserve some properties of the original data (e.g., the degree distribution in complex networks), the same trend observed from the null model captures its occurrence by chance. Then, by comparing the trend in the original data with that in the null model, the statistical significance of the observed trend in the original data can be measured. Table 3 shows the breakdown of ethnicity and gender of the ≈ 1 million users who shared URLs in Twitter between July and September 2016. We present a detailed description of the comparison with null models.

In this work, whenever we report the number of users with certain properties who share URLs on particular news media, we report Z-score by comparing the number of those users in the actual data with that in null models.

Consider that we are interested in users who are Asian and share BBC News. In this case, we denote by $|U_{BBC}|$ the number of users who share BBC News and $|U_{BBC}^{Asian}|$ by the number of Asian among them. To construct a null model, we create k random samples from a separate huge set of users, which is called Population, where each sample has exactly $|U_{BBC}|$ users. The demographic information of users in Population is inferred by *Face++*. For each sample, we count how many Asians are included, $|S_{BBC}^{Asian}|$. Then, the Z_{BBC}^{Asian} is computed as following:

$$Z_{BBC}^{Asian} = \frac{|U_{BBC}^{Asian}| - mean(|S_{BBC}^{Asian}|)}{std(|S_{BBC}^{Asian}|)} \quad (1)$$

where $mean(\cdot)$ is the mean and $std(\cdot)$ is the standard deviation of the values from multiple samples. 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, the size of Population is ≈ 1 million, and k=100.

3.8 Potential Limitations

There are a few limitations of our data, discussed next.

Accuracy of the inference by *Face++*. First, *(i)* we are limited by accuracy of *Face++* in the inference. *Face++* itself returns confidence levels for the inferred gender and race attributes and returns an error range for inferred age. In our data, the average confidence

level reported by *Face++* is 95.24 ± 0.020% for gender and 86.12 ± 0.032% for race, with a confidence interval of 95%. Besides that, as the performance of deep learning systems continues to improve, the inferred demographic attributes should become more accurate. Also, recent efforts have used *Face++* for similar tasks and reported high confidence in manual inspections of small samples [4, 44]; Another limitation, is that *(ii) Face++* reports race of recognizable faces from images but not the *ethnicity* (e.g. *Hispanic*); Finally, though *(iii)* we had discarded about 70% of the crawled users (i.e. those users whose profile pictures do not have a recognizable face or have profile pictures in which *Face++* recognized with low confidence). However, we note that the remaining final dataset is still representative and we only provide results that are statistically significant based on well known statistical tests.

Data. *(iv)* Our approach to identify users in U.S. may contain users located in the same time zone, but not in the U.S. We, however, believe that these users represent a small fraction of the users, given the predominance of active U.S. users in Twitter [14]; *(v)* We are using the 1% random sample off all tweets. Although the 1% random sample is not the best data to capture all the dynamics happening in Twitter, its limitations are known [28] and it is the best available option at our disposal.

Even with limitations, we believe that our dataset and methods can provide interesting insights on demographics and news sharing behaviors. In the following sections, we present and discuss the main results from characterizing news spreaders in Twitter along three dimensions: 1) their demographics (who they are), 2) their news shared (what they share), and 3) their impact (why they are important).

4 WHO ARE THE NEWS SPREADERS?

Our first research question is to understand who the spreaders are. We compare the demographics of news spreaders with 1) the spreaders of typical URLs in Twitter and 2) the Twitter followers of news media to see whether and to what extent they differ.

4.1 Typical URL Sharing Vs. News Sharing

Table 4 shows, for each news media, the proportion of news URL shares by different demographic groups. For example, for The New York Times, 54.1% of news shares are made by men and 79.2% of news shares are by Whites. The numbers in the parenthesis correspond to the Z-values, detailed in Section 3.7. We note that the Z-value indicates how news URL sharing behavior is similar or dissimilar from typical URL sharing behavior in terms of demographic composition.

By comparing between the news sources, we see some obvious patterns: 1) The Wall Street Journal is favored by Male (62.3%) more than Female (37.7%); 2) The New York Times has the most balanced gender distribution among spreaders (54.1% vs 45.9%); and 3) for The New York Times, The Guardian, and BBC News, the proportion of shares by Asians is greater than by AF-AM.

From a simple comparison to Table 2 which shows the demographic compositions of typical URL sharing behavior, we observed the following trends for all five news sources. First, Males share more news URLs than Female do. Male (54.84% of news spreaders) issue 54.1% to 62.3% of news URL shares. Secondly, Whites share more news URLs than other race groups–White (78.77% of total users) cover 78.2% to 82.2% of news URL shares.

The Z-values in Table 4 tell whether the differences between news spreaders and typical URL spreaders are statistically significant or not. The most strong tendency is observed for White-Male. White-Male share more news URLs than they share typical URLs and this tendency is strong ($Z > 11$[11]). Then, another observations is that White-Female are less likely to share news URLs than typical URLs ($Z < 0$) except for The New York Times. On average, White-Male make 46.8% and White-Female make 32.9% of news URL shares. From the two proportions, one may think this is because White-Female are less active than White-Male in Twitter. However, our method of comparing the news URL sharing behavior with typical

News media	Race (%)	Gender (%)		**Total:**
		Male	Female	
The New York Times	Asian	5.1 (-9.22)	5.9 (-18.02)	11.0 (-19.96)
	AF-AM	6.1 (-13.95)	3.7 (-15.01)	9.8 (-21.75)
	White	42.8 (26.24)	36.4 (2.86)	79.2 (31.32)
	Total:	54.1 (15.30)	45.9 (-15.30)	100.0
Reuters	Asian	3.6 (-8.06)	6.8 (-7.62)	10.4 (-12.09)
	AF-AM	7.3 (-3.02)	3.7 (-8.70)	10.9 (-9.03)
	White	47.0 (23.21)	31.7 (-4.89)	78.7 (16.38)
	Total:	57.9 (14.00)	42.1 (-14.00)	100.0
The Guardian	Asian	4.9 (-6.11)	5.9 (-9.75)	10.7 (-12.75)
	AF-AM	5.5 (-7.63)	3.3 (-9.77)	8.8 (-12.11)
	White	46.9 (23.03)	33.6 (-2.39)	80.5 (18.41)
	Total:	57.2 (13.24)	42.8 (-13.24)	100.0
The Wall Street Journal	Asian	4.9 (-3.91)	3.6 (-8.60)	8.5 (-9.43)
	AF-AM	6.1 (-3.41)	3.3 (-5.86)	9.4 (-6.68)
	White	51.3 (15.70)	30.8 (-3.35)	82.2 (12.23)
	Total:	62.3 (10.77)	37.7 (-10.77)	100.0
BBC News	Asian	5.3 (-2.64)	6.6 (-4.49)	12.0 (-5.11)
	AF-AM	7.1 (-1.91)	2.7 (-6.01)	9.8 (-5.76)
	White	46.2 (11.00)	32.1 (-2.36)	78.2 (8.04)
	Total:	58.6 (7.97)	41.4 (-7.97)	100.0

Table 4: Proportion of news shares by different demographic groups for each news source.

News media	Race (%)	Gender (%)		**Total:**
		Male	Female	
The New York Times	Asian	12.7 (6.69)	10.5 (0.28)	23.2 (5.28)
	AF-AM	11.4 (3.71)	3.9 (-4.35)	15.2 (-0.36)
	White	35.0 (2.41)	26.6 (-6.12)	61.5 (-4.08)
	Total:	59.1 (7.97)	40.9 (-7.97)	100.0
Reuters	Asian	11.3 (5.83)	7.9 (-2.97)	19.2 (1.71)
	AF-AM	16.9 (9.97)	3.6 (-4.64)	20.5 (3.98)
	White	39.5 (5.74)	20.8 (-10.31)	60.3 (-4.52)
	Total:	67.7 (15.81)	32.3 (-15.81)	100.0
The Guardian	Asian	8.5 (2.22)	7.8 (-3.34)	16.4 (-1.30)
	AF-AM	10.5 (2.79)	3.8 (-4.58)	14.3 (-1.04)
	White	41.4 (8.99)	27.9 (-5.82)	69.3 (1.80)
	Total:	60.4 (10.45)	39.6 (-10.45)	100.0
The Wall Street Journal	Asian	9.9 (4.13)	8.0 (-3.20)	17.9 (0.54)
	AF-AM	14.5 (8.55)	4.2 (-4.06)	18.8 (3.64)
	White	41.6 (6.97)	21.7 (-11.70)	63.3 (-3.28)
	Total:	66.0 (13.93)	34.0 (-13.93)	100.0
BBC News	Asian	12.5 (5.85)	11.3 (0.92)	23.8 (4.67)
	AF-AM	12.5 (4.58)	2.2 (-6.30)	14.7 (-0.59)
	White	34.6 (1.92)	26.9 (-5.13)	61.5 (-3.25)
	Total:	59.6 (7.57)	40.4 (-7.57)	100.0

Table 5: Proportion of distinct followers by different demographic groups for each news source.

[11] Z-value is minimum for BBC News, the largest Z-value is 26.24 for The New York Times.

URL sharing behavior can effectively tell that the difference is not because of the activity level, but of the type of URLs. White-Female do share a significant number of typical URLs.

4.2 Are Spreaders Similar to Followers of Media Sources?

In the previous analysis, we observed that White-Male are dominant in sharing news URLs. Then, would such pattern find for the Twitter followers of news sources?

Table 5 presents the demographics of Twitter followers of each news source. Again, the number in the parenthesis is Z-value, reporting how it differs from typical news sharing behavior. Compared to those users who share typical URLs, we observe two main differences of news media followers: 1) there are more male users ($Z > 0$); 2) except The Guardian, all the other four news sources have fewer White users ($Z < 0$). The New York Times and BBC News have more Asian followers and Reuters and The Wall Street Journal have more Asian and AF-AM users. This results in that the following three groups, Asian-Male, AF-AM-Male, and White-Male, are prominent in the followers of media sources ($Z > 0$). In addition, we observe that two news sources, The New York Times and BBC News, have positive Z-values for Asian Female followers.

For both type of users the followers and the spreaders we observe a "Male dominant" pattern, confirming that Male are more interested in news for consumption and spread. However, we find significant differences in demographic compositions between the followers and the spreaders of news. While the followers have a certain degree of racial equality, the spreaders are biased towards one particular race, White. This result is particularly important because so far it was known that individuals affiliated with news media play a large part in breaking the news [21]. Our observation indicates that breaking news is from not only those followers, but also from these news spreaders who are not necessarily following the news sources in Twitter.

5 WHAT NEWS SPREADERS SHARE

We study what news spreaders share along three distinct dimensions: the topical category of news, the demographic trait of the authors (journalist) of a news article, and the linguistic properties of news headlines.

5.1 By News Category

We firstly examine which categories of news are shared more by particular demographic groups. To this end, we standardized the names of topical categories for the analysis. For example, we grouped news categories relating to health and life and named "Health and Life" and grouped news categories relating to science and named "Science and Tech.".

Table 6 shows the proportion of news shares by each demographic group for each topical category. We consider only topics that were present in all news sources for this analysis. Foremost in Science and Tech, Business, and Politics, we can see the great gender differences. On average, 61.2% of news URLs of these three topics are shared by Male. In the others two categories, World and Health and Life, Female make more contributions (48.6% of shares).

Category	Race (%)	Gender (%)		**Total:**
		Male	Female	
World	Asian	4.3 (-6.96)	7.8 (-6.32)	12.1 (-9.84)
	AF-AM	6.1 (-6.47)	3.2 (-9.35)	9.3 (-12.53)
	White	40.4 (13.71)	38.3 (4.62)	78.6 (17.67)
	Total:	50.8 (4.55)	49.2 (-4.55)	100.0
Health and Life	Asian	6.8 (-0.18)	7.0 (-2.76)	13.8 (-2.25)
	AF-AM	3.3 (-3.82)	3.7 (-2.94)	7.0 (-5.54)
	White	41.9 (4.77)	37.3 (0.93)	79.2 (5.97)
	Total:	52.0 (1.89)	48.0 (-1.89)	100.0
Science and Tech	Asian	5.2 (-3.55)	5.4 (-6.98)	10.5 (-7.61)
	AF-AM	6.3 (-3.25)	1.7 (-10.12)	8.0 (-9.17)
	White	52.6 (19.17)	28.8 (-5.95)	81.4 (12.34)
	Total:	64.1 (15.74)	35.9 (-15.74)	100.0
Business	Asian	4.0 (-4.95)	5.3 (-6.93)	9.3 (-8.13)
	AF-AM	7.0 (-2.54)	3.1 (-5.69)	10.0 (-6.30)
	White	49.4 (15.59)	31.3 (-3.48)	80.7 (10.45)
	Total:	60.3 (9.76)	39.7 (-9.76)	100.0
Politics	Asian	5.5 (-3.06)	4.6 (-9.52)	10.1 (-9.94)
	AF-AM	6.3 (-4.01)	3.2 (-7.53)	9.5 (-8.18)
	White	47.3 (16.91)	33.1 (-2.58)	80.4 (14.20)
	Total:	59.1 (13.23)	40.9 (-13.23)	100.0

Table 6: Number of shares by category.

When compared to typical URL sharing behavior, we observe the tendency of White-Male sharing news URLs for all categories ($Z > 0$), but the tendency is stronger for Science and Tech, Business, and Politics ($Z > 9.76$) than World ($Z = 4.55$) and Health and Life ($Z = 1.89$). One interesting observation is that White-Female do share more news URLs of World and Health and Life categories than the typical URLs ($Z > 0$).

To understand better how demographic traits relate to topical preferences, we compute the relative preferences of each demographic group to ten topical categories (see Figure 1). News articles about Tech are more likely to be shared by Male than Female. We then see White are more likely to share news about Health and Tech while Asian and AF-AM participate more in sharing news about Sports and Arts. Lastly, Science is favored by Asian but Business is favored by AF-AM. Our analysis shows that demographic groups have different topical tastes in sharing. This guides us how news media publish their contents to target appropriate user segments.

5.2 By Author's Demographics

In this section, we study how the gender of a journalist who wrote a news article influences its shares. While some differences in topics written [26] or sources used [45] between male and female journalists have been reported [26] , its appealing to each demographic group has not been fully explored.

Table 7 shows the demographics of the authors for each news source. Overall, the proportion of Male authors are higher than that of Female authors–on average, 60.04% of the authors are Male. Reuters and BBC News have more skewed gender distributions than the other three sources. In terms of race, most of the authors are White (83.8% on average across five media sources), followed by Asian authors (10.5%). We observe only 5.7% of the authors are AF-AM and strikingly low fraction of AF-AM Female authors (1.42%).

Table 8 shows the proportion of the spreaders who shared any news URLs written by a certain author demographic group for each news source.

News Media	Race (%)	Gender (%)		Total:
		Male	Female	
The New York Times	Asian	4.9	5.8	10.7
	AF-AM	3.9	0.9	4.8
	White	49.4	35.1	84.5
	Total:	58.1	41.9	100.0
Reuters	Asian	6.8	6.0	12.8
	AF-AM	4.3	2.3	6.6
	White	51.3	29.3	80.6
	Total:	62.5	37.5	100.0
The Guardian	Asian	3.4	4.6	8.1
	AF-AM	3.8	1.2	5.0
	White	50.4	36.6	87.0
	Total:	57.6	42.4	100.0
The Wall Street Journal	Asian	7.0	6.1	13.1
	AF-AM	2.9	1.6	4.5
	White	47.9	34.5	82.4
	Total:	57.8	42.2	100.0
BBC News	Asian	3.2	4.7	7.9
	AF-AM	6.3	1.1	7.4
	White	54.7	30.0	84.7
	Total:	64.2	35.8	100.0

Table 7: Demographic characteristics of the collected authors by news source.

5.2.1 Author's Gender. Does the gender of an author affect the spreading behavior? For The New York Times and Reuters, the proportion of Male spreaders is not significantly different (< 2%) no matter the gender of the author is. However, in the rest three others sources, Male tend to share more news URLs written by Male–the difference is 12.4% for BBC News, 7.4% for The Wall Street Journal, and 5.3% for The Guardian. While the effect of the gender of the authors on spreading behavior exists, this might be a mere effect of biological differences in topical tastes–Male and Female journalists write only the topics that readers of the same gender are interested in.

To control the effect of the topics, we use a Chi-square test [9] to find which topics are written significantly more by Female (or Male) journalists and which topics are significantly more shared by Female (or Male) spreaders. Table 9 shows the graphical presentation of the statistically significant results by Chi-square test statistics ($p < 0.05$). In the table, an upward pointing arrow represents a higher tendency in writing or sharing. For example, Male authors write news significantly more about Sport and Opinion, and Female authors write about Health. There are no topics that authors and spreaders have the same gender differences except for Health. Therefore, the gender difference in spreading behavior is unlikely driven by that in journalists' choice of the topics. We bring the potential explanation in later section based on linguistic component of news.

5.2.2 Author's Race. Does the race of an author affect the spreading behavior? We observe that the proportion of Asian spreaders are significantly difference across different race of the authors in all news sources except The New York Times. For Reuters, The Guardian, and The Wall Street Journal, Asian spreaders are more likely to share news URLs written by Asian or AF-AM authors. Compared to the proportion of shares by Asian (Table 4) which are 10.4%, 10.7%, and 8.5% for those three news sources, respectively, the proportion of the news URLs shares written by Asian authors are increased by 26.9%, 23.4%, and 47.1%, respectively. For AF-AM users, we did not find the same pattern. Lastly, BBC News has a

(a) The New York Times

Authors (%)	Spreaders (%)		
	Male	Female	
Male	54.8	45.2	
Female	53.1	46.9	
	Asian	AF-AM	White
Asian	11.0	10.7	78.3
AF-AM	11.3	11.6	77.1
White	10.7	10.3	79.1

(b) Reuters

Authors (%)	Spreaders (%)		
	Male	Female	
Male	58.7	41.3	
Female	57.5	42.5	
	Asian	AF-AM	White
Asian	13.2	10.4	76.5
AF-AM	14.0	9.6	76.3
White	9.5	10.9	79.6

(c) The Guardian

Authors (%)	Spreaders (%)		
	Male	Female	
Male	59.7	40.3	
Female	54.4	45.6	
	Asian	AF-AM	White
Asian	13.2	10.0	76.8
AF-AM	14.1	11.1	74.8
White	10.5	9.7	79.8

(d) The Wall Street Journal

Authors (%)	Spreaders (%)		
	Male	Female	
Male	66.9	33.1	
Female	59.6	40.4	
	Asian	AF-AM	White
Asian	12.5	9.2	78.4
AF-AM	12.3	9.9	77.8
White	9.3	9.8	80.9

(e) BBC News

Authors (%)	Spreaders (%)		
	Male	Female	
Male	62.4	37.6	
Female	50.0	50.0	
	Asian	AF-AM	White
Asian	3.4	6.9	89.7
AF-AM	13.3	40.0	46.7
White	11.3	9.1	79.6

Table 8: Confusion matrixes for news authors and spreaders by news source.

(a) Gender

Topic	Author		Spreader	
	Female	Male	Female	Male
Sport	↓	↑		
Opinion	↓	↑		
Health	↑	↓	↑	↓
Tech			↓	↑
Business			↓	↑

(b) Race

Topic	Author			Spreader		
	Asian	AF-AM	White	Asian	AF-AM	White
World	↑		↓			
Tech	↑		↓			
Opinion	↓		↑			
Sports				↑	↑	↓
Art				↑		↓

Table 9: Discriminative topics for gender and race groups by authors and spreaders.

strong tendency that AF-AM share extensively news URLs written by AF-AM and Asian.

Table 9(b) shows the discriminative topics for each racial group of authors and spreaders. Asian authors are writing more about World and Tech than White. White authors write more opinionated news articles than Asian. For spreaders, Asian and AF-AM share more Sports news than White. News about Arts is favored by Asian more than White. Once again, we do not find any relationship between the topical interests of a certain racial author group and those of a certain racial spreaders group.

5.3 By LIWC Analysis

Linguistic Inquiry and Word Count (LIWC) [33] is a dictionary-based text mining software. Since it has been proposed, it has been widely used for a number of different tasks, including sentiment analysis [35] and discourse characterization in social media platforms [15]. Next, we use LIWC to characterize differences in the content shared by different demographic groups. Its latest version, LIWC 2015 (used in this work), defines about 90 linguistic categories and classifies more than 6,400 words into those categories [32]. For example, the word 'cried' falls into the sadness, negative emotion, overall affect, verbs, and past focus categories. Then, in a given text, the LIWC software finds the occurrence of the words in each category. The output is the proportion of the words in each category to the total words in the text.

Table 10 presents the result of LIWC analysis of headlines shared by Male spreaders and Female spreaders. For comparison purposes, we also show the result of effects of gender on language use [30]. We show only LIWC dimensions that have more than 20% differences between Male and Female and omit the rest because the number of the whole dimensions is more than 90.

In our data, we find exactly the same trend as [30]: Female share headlines including more first-person singular pronouns, third-person pronouns, negations, words about ingestion (e.g., dish, eat, or pizza), assent (e.g., agree, yes, or ok), and female references (e.g., girl, her, or mom), and Male share headlines including more

LIWC Dimension	Our data	Newman et al. [30]
Pronouns		
First-person singular	M<F	M<F
Third-person	M<F	M<F
Linguistic dimensions		
Negations	M<F	M<F
Current concerns		
Money	M>F	M>F
Biological process		
Ingestion	M<F	-
Spoken categories		
Assent	M<F	-
Swear words	M>F	M>F
Female references	M<F	-

Table 10: LIWC analysis of ours and [30].

words about money (e.g., audit, cash, or owe), and swear words (e.g., damn, or shit). Considering that [30] observed those language usage patterns in the texts Male or Female *write*, finding the same patterns in the texts he or she *shares* is surprising and interesting. The spreaders are likely to share the news that is aligned with the language usage of their own. While many research have focused on attracting more clicks by tweaking headlines, such as including named-entities in headlines [23], we show that those studies can be extended to target specific user segments.

In addition, we find some results that are aligned with some stereotypes of races (e.g. Asian share headlines including more words related to family). However, we omit the result of LIWC by race of the spreaders because there have been no available references for a systematic comparison.

6 IMPORTANCE OF SPREADERS

Finally, we study the impact of understanding news spreaders in two ways: 1) extended readership by news spreaders and 2) understanding news popularity and demographics of news spreaders.

For the first, we compare the original followers and followers of spreaders by the number and the demographics. That is, we analyzed how spreaders extend news media's readers. For example, if followers of the The New York Times are usually white male but spreaders of The New York Times URLs have a lot of Asian followers, then, the role of spreaders is really important not only because it increases the number of audience but also because it brings "different" audiences. The results are shown below in detail.

6.1 Extended Readership by Spreaders

Ideally, to study the audience size reached by spreaders that is not reached directly by news sources profiles, 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 these users surpasses a billion users, which is unfeasible to be crawled given our resources. As an attempt to provide evidence that spreaders can largely benefit audience of news papers in social media systems, Table 11 contrasts the number of followers of the news media profiles and the sum of the number of followers of the spreaders of each news source. Although these results do not quantify exactly the extent to which spreaders are able to increase the audience size

News Media	#Followers (news media)	# Followers (spreaders)
The New York Times	32,626,611	67,458,732
Reuters	15,946,449	11,119,453
The Guardian	6,154,465	21,120,210
The Wall Street Journal	12,563,525	6,193,775
BBC News	27,871,624	4,713,614

Table 11: Total/Real number of followers of the news sources in Twitter and number of followers of the spreaders that shared news of the news source.

of news sources, it clearly shows that they play a very important role in many news source audiences. For example, the number of followers in our sample of spreaders from NYTimes contains more than double the number of followers of The New York Times.

We move onto demographic of the followers of news spreaders. First, we collected followers from a sample of 25% of spreaders from our dataset. For this data sample, the average confidence level for the number of the followers of the spreaders is 6111.154 ± 66396.94, with a confidence interval of 95%. After that, we analyze the demographic characteristics of the followers of the spreaders.

Table 12 shows the demographics of the followers of news spreaders. Compared with the demographics of the followers of news sources (Table 5), we observe the increase in the percentage of Female–the average increase is 9%. Besides that, for race, the percentage of White is higher–the average increase is 16%. We tried to test whether this difference in demographics of spreaders' followers and those of the original followers is statistically significant. We define the demographic distribution of the audience for each news media as a six-long vector whose element is a proportion of each demographic group (e.g., Male-Asian, Female-Asian, ..., and Female-White), respectively. With these vectors, we use the Kolmogorov-Smirnov test, which is a widely used statistical test to check whether two distributions are generated from an identical reference distribution. However, the difference is not statistically significant (for The New York Times, D = 0.5, p-value = 0.1641). The main reason is that the length of the vector, six, is too short to get statistical evidence. In future work, we will build demographic vectors for multiple snapshots and compute the statistical significance by concatenating those vectors.

6.2 News Popularity and Demographics

In the previous section, we show that understanding news spreaders is important as they extend the readership of news media. Another important aspect is whether the demographic traits of news spreaders are relating to the popularity of news. To this end, we collect the number clicks for each news URL using the Bit.ly API[12]. Then, we compare the popularity of news articles shared by different demographic groups to know whether a certain demographic group share news URLs likely to be more popular.

For gender group, we observe that the news items shared by Female are more clicked that those shared by Male. The differences are statistically significant by Kruskal-Wallis H-test ($H = 7.719, p < 0.005$). For race, the news articles shared by Asians are more clicked

[12]https://dev.bitly.com/

News Media	Race (%)	Gender (%)		**Total:**
		Male	Female	
The New York Times	Asian	4.8	5.6	10.4
	AF-AM	6.3	4.2	10.5
	White	41.5	37.5	79.1
	Total:	52.7	47.3	100.0
Reuters	Asian	4.8	5.4	10.2
	AF-AM	6.3	4.0	10.4
	White	42.3	37.1	79.4
	Total:	53.4	46.6	100.0
The Guardian	Asian	4.8	5.3	10.1
	AF-AM	6.1	3.8	9.9
	White	42.7	37.2	80.0
	Total:	53.6	46.4	100.0
The Wall Street Journal	Asian	4.8	5.3	10.1
	AF-AM	6.1	3.9	10.0
	White	43.0	36.9	79.9
	Total:	54.0	46.0	100.0
BBC News	Asian	4.8	5.3	10.1
	AF-AM	6.1	3.8	9.8
	White	42.9	37.2	80.1
	Total:	53.7	46.3	100.0

Table 12: Demographic characteristics of each the followers of the spreaders by news source.

($H = 6.659, p < 0.005$). The results show that the demographic information of news spreaders can potentially help in predicting the popularity of news articles.

7 CONCLUDING DISCUSSION

The increasing diffusion of news in social media systems, associated with the great power provided to users along the dissemination process, are making these platforms a fertile ground for misleading or fake news propagation. The growing use of Twitter as a news' channel highlights the importance of characterizing news spreaders to understand who they are, what they share and their impact. Next, we briefly discuss implications of our main findings and discuss directions we aim to explore next.

Bias on breaking news stories: A widely used tool that users use to find breaking news-stories in online social networks is the Trending stories (or topics) [19, 40]. Recently, Facebook has been involved in many controversies related to trending stories [18]. First, Facebook involved human curators as part of its process to identify trending stories. A main criticism was that human curators could bias the final list of stories. Then, Facebook removed the human intervention and followed the popular perception that data-driven algorithms would not be biased as they simply process data. Our results, however, shows the data itself is biased, at least in terms of the demographic groups considered. We show that demographic groups of white and male users tend to share more news in Twitter. Our results also quantify the existing bias on Twitter shares towards specific demographic groups across news categories and other dimensions. Thus, our work contributes with a new and important perspective to the emerging debate in the community centered around concerns about bias and transparency of decisions taken by algorithms operating over user-generated data. Finally, we believe that the increasing availability of information about demographics will help the development of systems that promote more diversity and less inequality to users. Thus, as a final contribution of our effort, we intend to release our demographic dataset to the research

community by the time of publication of this study.

Personalized news recommendations: Our analysis shows different user behaviors in terms of news sharing and also highlight demographic differents in terms of user interests. Identifying intrinsic characteristics of the users who spread the news in the online world and identifying how users interest across demographics is a key step towards the development of a framework that can promote the customization of the user experience using social media for news digest. We aim at further exploring this topic as part of our future work by investigating the discriminative power of demographic, linguistic, and network features in predicting a user's interest in specific news and news topics.

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, Meeyoung Cha, Krishna Gummadi, and Jon Crowcroft. 2011. Media landscape in Twitter: A world of new conventions and political diversity. In *ICWSM*.
[2] Jisun An and Haewoon Kwak. 2016. Multidimensional Analysis of Gender and Age Differences in News Consumption. In *Computation + Journalism Symposium*.
[3] Jisun An and Haewoon Kwak. 2017. What Gets Media Attention and How Media Attention Evolves Over Time - Large-scale Empirical Evidence from 196 Countries. In *ICWSM*.
[4] Jisun An and Ingmar Weber. 2016. # greysanatomy vs.# yankees: Demographics and Hashtag Use on Twitter. In *ICWSM*.
[5] Cory L Armstrong and Fangfang Gao. 2011. Gender, Twitter and news content: An examination across platforms and coverage areas. *Journalism Studies* 12, 4 (2011), 490–505.
[6] Cameron Blevins and Lincoln Mullen. 2015. Jane, John... Leslie? a historical method for algorithmic gender prediction. *Digital Humanities Quarterly* 9, 3 (2015).
[7] Jonathan Bright. 2016. The Social News Gap: How News Reading and News Sharing Diverge. *Journal of Communication* 66, 3 (2016), 343–365.
[8] John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *EMNLP*.
[9] George Casella and Roger L Berger. 2002. *Statistical inference*. Vol. 2. Duxbury Pacific Grove, CA.
[10] Pew Research Center. 2016. News Use Across Social Media Platforms 2016. (2016).
[11] Meeyoung Cha, Fabrício Benevenuto, Hamed Haddadi, and Krishna Gummadi. 2012. The world of connections and information flow in twitter. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 42, 4 (2012), 991–998.
[12] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *ICWSM*.
[13] 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 *ICWSM*.
[14] Alex Cheng, Mark Evans, and Harshdeep Singh. 2009. Inside Twitter: An in-depth look inside the Twitter world. *Report of Sysomos, June, Toronto, Canada* (2009).
[15] 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 *ICWSM*.
[16] Evandro Cunha, Gabriel Magno, Virgilio Almeida, Marcos André Gonçalves, and Fabricio Benevenuto. 2012. A gender based study of tagging behavior in twitter. In *HT*.
[17] 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 *CSCW*.
[18] Caitlin Dewey. 2016. What we really see when Facebook Trending picks stories for us. washingtonpost.com/news/ the-intersect/wp/2016/05/20/what-we-really-see-when- facebook-trending-picks-stories-for-us. (May 2016).
[19] Facebook. 2016. Search FYI: An Update to Trending. newsroom.fb.com/news/2016/08/search-fyi-an-update-to-trending/. (2016).
[20] Eric Gilbert, Saeideh Bakhshi, Shuo Chang, and Loren Terveen. 2013. I need to try this?: a statistical overview of pinterest. In *SIGCHI*.
[21] Mengdie Hu, Shixia Liu, Furu Wei, Yingcai Wu, John Stasko, and Kwan-Liu Ma. 2012. Breaking news on twitter. In *SIGCHI*.
[22] 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 *WWW(Companion Volume)*.
[23] Joon Hee Kim, Amin Mantrach, Alejandro Jaimes, and Alice Oh. 2016. How to Compete Online for News Audience: Modeling Words that Attract Clicks. In *SIGKDD*.
[24] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *WWW*.
[25] 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: Analyzing Microtext*, Vol. 13.
[26] Dianne Lynch. 1993. Catch 22?: Washington Newswomen and Their News Sources. *Newspaper Research Journal* 14, 3-4 (1993), 82–92.
[27] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. Understanding the Demographics of Twitter Users.. In *ICWSM*.
[28] Fred Morstatter, Jürgen Pfeffer, and Huan Liu. 2014. When is it biased?: assessing the representativeness of twitter's streaming API. In *WWW*.
[29] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. 2011. Bad news travel fast: A content-based analysis of interestingness on twitter. In *WebScience*.
[30] Matthew L Newman, Carla J Groom, Lori D Handelman, and James W Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes* 45, 3 (2008), 211–236.
[31] 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 *ICWSM*.
[32] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.
[33] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71 (2001).
[34] Julio Reis, Fabrıcio Benevenuto, Pedro OS de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. 2015. Breaking the news: First impressions matter on online news. In *ICWSM*.
[35] 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).
[36] Tiago Rodrigues, Fabrício Benevenuto, Meeyoung Cha, Krishna P. Gummadi, and Virgílio Almeida. 2011. On Word-of-Mouth Based Discovery of the Web. In *SIGCOMM(IMC)*.
[37] Daniel M Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A Huberman. 2011. Influence and passivity in social media. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.
[38] Dhavan V Shah, Joseph N Cappella, W Russell Neuman, Stuart Soroka, Lori Young, and Meital Balmas. 2015. Bad news or mad news? Sentiment scoring of negativity, fear, and anger in news content. *The ANNALS of the American Academy of Political and Social Science* 659, 1 (2015), 108–121.
[39] 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 (2015).
[40] Twitter. 2010. To Trend or Not to Trend. blog.twitter.com/2010/to-trend-or-not-to-trend. (2010).
[41] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's wikipedia? Assessing gender inequality in an online encyclopedia. In *ICWSM*.
[42] Shaomei Wu, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Who says what to whom on twitter. In *WWW*.
[43] Qi Yin, Zhimin Cao, Yuning Jiang, and Haoqiang Fan. 2016. Learning deep face representation. (July 26 2016). US Patent 9,400,919.
[44] Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, and others. 2014. Inferring international and internal migration patterns from twitter data. In *WWW*.
[45] Geri Alumit Zeldes, Frederick Fico, and Arvind Diddi. 2007. Race and gender: An analysis of the sources and reporters in local television coverage of the 2002 Michigan gubernatorial campaign. *Mass Communication & Society* 10, 3 (2007), 345–363.