Biases on Social Media Data

(Keynote Extended Abstract)

Ricardo Baeza-Yates

rbaeza@acm.org

NTENT &

Khoury College of Computer Sciences

Northeastern University at Silicon Valley

CA, USA

ABSTRACT

Social media data is often used to pulse the opinion of online communities, either by predicting sentiment or stances (*e.g.*, *political*), to mention just two typical use cases. However, those analysis assume that the data samples really represent the underlying demographics of the overall community, both, in number and characteristics, which in most cases is not true. As a result, extrapolating these results to larger populations usually do not work. This happens because social media data is inherently biased, mainly due to two facts: (1) not all people is equally active in social media platforms and most of them are really passive; and (2) there are demographic biases in gender and age, among other attributes. Hence, the questions of how representative is the data and if is possible to make it representative are of crucial importance. We also discuss related issues such as using public samples of mostly private platforms as well as typical errors in the analysis of such data.

CCS CONCEPTS

• **Networks** → **Social media networks**; • **Social and professional topics** → **User characteristics**; • **Computing methodologies** → *Machine learning*.

KEYWORDS

Bias, representativeness, data samples, gender and age prediction.

ACM Reference Format:

Ricardo Baeza-Yates. 2020. Biases on Social Media Data: (Keynote Extended Abstract). In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion)*, April 20–24, 2020, Taipei, Taiwan. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3366424.3383564>

SUMMARY

Social media is the Internet projection of our cultures through the Web and mobile apps. A kind of enlarging mirror of the good and bad parts of ourselves, but also a tricky mirror [6]. Hence, extracting unbiased information from social media data is a non-trivial problem. Below we summarize the main challenges and discuss some solutions as well as misuses.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

© 2020 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.

ACM ISBN 978-1-4503-7024-0/20/04.

<https://doi.org/10.1145/3366424.3383564>

Inequality of Access

The first challenge is that Internet penetration is very heterogeneous (from 98% in Iceland to 2% in South Sudan) and reaches just 59% of the world population.¹ This mainly leaves out developing countries as well as poor people everywhere, creating an immediate wealth bias in Internet as a data source. Indeed, the penetration does not reach 40% in Africa while goes up to almost 95% in North America. In contrast, mobile phone penetration is larger than Internet penetration, specially in developing countries, reaching almost 67% of the Earth population.

Social Media and Data Privacy

Social media or online social networks (OSN) users are estimated in 3.8 billions, which represent almost 49% of the world population and 83% of the people with Internet access. In Table 1 we show the main online social networks, including implicit ones (chat apps) and excluding the ones focused in China (QQ, Qzone, Sina Weibo, and Kuaishou) with the exception of WeChat. These numbers must be considered as proxies to the real numbers as they come from different sources where not always there is consensus.

As we can see in Table 1, in many cases data is not public, hence we need to assume that people that makes their data public is similar to the one the keeps their data private. This is a very strong assumption and there are no studies, to the best of our knowledge, showing this (can only be done by each OSN). In addition, in many cases we sample the data with an API provided by the OSN. Hence we need to also assume that this sample is random, but probably is not completely true as at least some filtering might be done (*e.g.*, adult or hate content).

Biases

The first bias is *activity* bias [2]. That is, not all people are active all the time. Indeed, Nielsen proposed the 90-9-1 rule to convey participation inequality in Internet [5]. That is, in any segment of time, 90% of the people is passive (lurkers), 1% is heavily active and 9% react to that activity. For example, in Table 1 we show active monthly and daily users when available, but the overall number of users of each OSN is in many cases much higher (*e.g.*, for LinkedIn is about 670 million). If we consider the percentage of active users that generate half the content, we have found 7% and 2%, for Facebook and Twitter, respectively [1]. These values are consistent with the 90-9-1 rule and imply that content per user is quite heterogeneous, not only in content, but also in volume.

¹<https://www.internetworldstats.com/stats.htm>

OSN	Public data?	Active users (M)		Gender (%)		Age (%)		Worldwide penetration	Geographic coverage
		Monthly	Daily	Female	Male	18-34	35-65		
Facebook	Some	2450	1650	54	46	65		31%	Worldwide but China & others
YouTube	Mostly	2000	150	38	62	✓		26%	Worldwide but China & others
Whatsapp	None	1600	1000	50	50			21%	Worldwide but less in USA
FB Messenger	None	1300		50	50			17%	Worldwide but China & others
WeChat	None	1150		50	50			15%	Mainly China
Instagram	Mostly	1000	600	56	44	65		13%	Worldwide
Tik Tok	Mostly	800		44	56	>41		10%	Worldwide
Reddit	All	430		35	65			6%	Worldwide
Snapchat	Some	380	135	61	39	60		5%	Partial
Twitter	All	340	150	34	66		63	4%	Worldwide
Pinterest	All	330		70	30		✓	4%	Worldwide
LinkedIn	All	310		43	57	> 37		4%	Worldwide

Table 1: Main online social networks (some are *implicit*) and their characteristics (age ranges are partial).
(Sources: statista.com, hootsuite.com, dustinsout.com, khoros.com, wikipedia.org)

Figure 1: Age bias in a Twitter user sample for Chile.

The second bias is demographic. Gender, age, and other characteristics do not necessarily reflect the general population. As we can see in Table 1, some social media are dominated by women while others are dominated by men (the numbers represent an informed guess as in some cases there is no consensus among the different sources). Similarly, some OSNs are dominated by young people while others are dominated by adults over 35 years old.

Discussion

Demographic bias can be mitigated by segmenting the users by gender and age ranges. To do this we need to build classifiers for gender and age range prediction, using users that self-report those attributes as training data. Then, we must compare the percentage of users in each segment with the latest census of the country where we are doing the analysis, to know how to weight each segment so we get the real representation of it. We did this in an analysis of Twitter data during the legislation of a controversial abortion law in Chile and when compared to real surveys, we got representative results [4]. Figure 1 shows the differences in age segments for self-reported, predicted and overall population. In this same work we estimated that 43% of the users were women, much more than the world average, probably due to the topic under study. Hence, demographic biases can be very different from one country to another.

Finally, when analyzing social media data, there is the tendency to look at the top- k influencers. For example, which percentage of them have a property (e.g., are foreign). However, this is completely

bogus as the percentage depends on the value of k . In fact, we can choose k to maximize this percentage which will be much larger than the real percentage when all users are considered (a number much larger than k). We have seen this mistake in many studies, including some that are well publicized [3].

BIOGRAPHY

Ricardo Baeza-Yates is, since June 2016, CTO of NTENT, a semantic search technology company based in California, USA. He is also the (part-time) Director of Data Science Programs at Northeastern University, Silicon Valley campus, since August 2017. Before, he was VP of Research at Yahoo Labs, based in Barcelona, Spain, and later in Sunnyvale, California, from January 2006 to February 2016. He is co-author of the best-seller Modern Information Retrieval textbook published by Addison-Wesley in 1999 and 2011 (2nd ed), that won the ASIST 2012 Book of the Year award. From 2002 to 2004 he was elected to the Board of Governors of the IEEE Computer Society and between 2012 and 2016 was elected for the ACM Council. Since 2010 is a founding member of the Chilean Academy of Engineering. In 2009 he was named ACM Fellow and in 2011 IEEE Fellow, among other awards and distinctions. He obtained a Ph.D. in CS from the University of Waterloo, Canada, in 1989, and his areas of expertise are web search and data mining, information retrieval, data science and algorithms in general.

REFERENCES

- [1] Ricardo Baeza-Yates and Diego Saez-Trumper. 2015. Wisdom of the Crowd or Wisdom of a Few? An Analysis of Users' Content Generation. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media (HT '15)*. ACM, New York, NY, USA, 69–74.
- [2] Ricardo Baeza-Yates. 2018. Bias on the Web. *Communications of ACM* 61, 6 (June 2018), 54–61.
- [3] Ricardo Baeza-Yates. 2020. The Amazing Errors of Alto Analytics' Analysis of Chilean Social Networks (in Spanish). Medium, https://medium.com/@rbaeza_yates/los-asombrosos-errores-del-2b0225c2e622.
- [4] Eduardo Graells-Garrido, Ricardo Baeza-Yates, and Mounia Lalmas. 2019. How Representative is an Abortion Debate on Twitter? In *Proceedings of the 10th ACM Conference on Web Science (WebSci '19)*. ACM, New York, NY, USA, 133–134.
- [5] Jakob Nielsen. 2006. The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities. <https://www.nngroup.com/articles/participation-inequality/>.
- [6] Jia Tolentino. 2019. Trick Mirror: Reflections on Self-Delusion. *Random House*, NY, USA.