J. Mantas et al. (Eds.)

© 2020 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).

doi:10.3233/SHTI200484

Application of Topic Modeling to Tweets as the Foundation for Health Disparity Research for COVID-19

Michelle ODLUM [a], Hwayoung CHO [b], Peter BROADWELL [c], Nicole DAVIS [d], Maria PATRAO [e], Deborah SCHAUER, Michael E. BALES [f], Carmela ALCANTARA [g] and Sunmoo YOON [h,1]
[a] *School of Nursing, Columbia University, USA*
[b] *College of Nursing, University of Florida, USA*
[c] *Stanford University, USA*
[d] *School of Nursing, Clemson University, USA*
[e] *New York Presbyterian Hospital, USA*
[f] *Weill Cornell Medicine, USA*
[g] *School of Social Work, Columbia University, USA*
[h] *Columbia University Irving Medical Center, USA*

Abstract. We randomly extracted publicly available Tweets mentioning COVID-19 related terms (n=2,558,474 Tweets) from Tweet corpora collected daily using an API from Jan 21st to May 3rd, 2020. We applied a clustering algorithm to publicly available Tweets authored by African Americans (n=1,763) to detect topics and sentiment applying natural language processing (NLP). We visualized fifteen topics (four themes) using network diagrams (Newman modularity 0.74). Compared to the COVID-19 related Tweets authored by others, positive sentiments, cohesively encouraging online discussions (e.g., Black strong 27.1%, growing up Blacks 22.8%, support Black business 17.0%, how to build resilience 7.8%), and COVID - 19 prevention behaviors (e.g., masks 4.7%, encouraging social distancing 9.4%) were uniquely observed in African American Twitter communities. Application of topic modeling techniques to streaming social media Twitter provides the foundation for research team insights regarding information and future virtual based intervention and social media based health disparity research for COVID-19.

Keywords. social media, pandemic, virtual intervention, health disparities

1. Introduction

To date, a total of 3,525,116 confirmed COVID-19 cases and 243,540 deaths have been reported worldwide. While there are several ongoing clinical trials, there are no medicines that have been shown to prevent or cure the disease. Notably, African Americans are disproportionately affected by Covid-19 [1]. Topic modeling applies statistical analyses to counts of words or word groups that co-occur within text documents in order to reveal latent word groupings with potential semantic significance

[1] Corresponding Author, Sunmoo Yoon, PhD, MS, RN, 630W 168 street, PH105, New York, NY, 10032, USA; E-mail: sy2102@cumc.columbia.edu.

across a collection of documents. Compared to manual reading and keyword/key phrase extraction algorithms (TF-IDF, TextRank), topic modeling offers the ability to find semantically meaningful groupings of multiple words across very large document collections, even in cases when all of the words do not appear together within any single document or section [2]. Topic modeling is typically applied in health science to facilitate meta-analyses of large collections of research publications, or for generating high-level summaries of discussions (e.g., on social media) among groups of interest. The purpose of the study was to apply topic modeling techniques to Tweets as the foundation to inform designs of culturally sensitive interventions targeting African Americans on virtual platforms for COVID-19.

2. Methods

We applied topic modeling and sentiment analysis for content mining of publicly available Tweet corpus mentioning COVID-19 (keywords: covid19, Covid, covid-19, COVID, corona, coronavirus). We used NCapture, ORA and a Terremoto High Performance Computing Cluster (https://cuit.columbia.edu/shared-research-computing-facility). We randomly extracted publicly available Tweets mentioning COVID-19 related terms (n=2,558,474 Tweets) from Tweet corpora collected daily using an API (https://osf.io/qruf3/?view_only=08302960d9f4486083e82971a1eda125) from Jan 21st to May 3rd, 2020. First, 327 XML files were merged using Perl regular expression commands and shell scripting on the Terremoto cluster. Second, we randomly extracted Covid-19 related Tweets using publicly open African American Twitter community hashtags (#blacktwitter, #staywoke, #blacklivesmatter). Third, we applied NLP and a clustering Newman algorithm to the publicly available Tweet corpus. The relative frequency of 55,035 n-grams from both corpora were compared using Chi-square in order to identify n-grams uniquely occurring in Tweets authored by public African American Twitter communities. Next, we visualized 15 topics using a network diagram (Newman modularity 0.74). Lastly, experts in African American culture studies, public health, nursing and behavioral science performed qualitative thematic analysis on the detected topics. The larger study was approved by the Institutional Review Board (IRB).

3. Results

Positive sentiments, cohesively encouraging online discussions (e.g., Black strong 27.1%, growing up black 22.8%, support black business 17.0%, how to build resilience 7.8%) , and COVID-19 prevention behaviors (e.g., masks 4.7%, encouraging social distancing 9.4%) were uniquely observed in Tweet corpus authored by African Americans. Top 50 unique n-grams in COVID-19 related Tweets authored by African Americans and their frequencies are summarized in **Table 1** (χ^2 p-value < 2.2e-16).

Table 1. Top unique n-grams in COVID-19 related public Tweets within African American Communities

n gram	frequency	n gram	frequency
African Americans	40.72%	stylist	6.26%
xcellence	32.66%	Monday motivation	6.04%
black strong	27.07%	woo young (k-pop singer)	6.04%
growing up black	22.82%	calling	5.82%
support black business	17.00%	quarantine	5.82%
warriors	16.11%	hospitals	5.59%
died (deaths, dying)	13.20%	mortalities	5.37%

doctors	12.30%	patients	5.15%
lock down hustle	11.63%	wards	5.15%
black girl magic	10.07%	blue lives matter	4.92%
cases	10.07%	essential not expendable	4.92%
Michigan protest	9.84%	gentiles	4.92%
communities	9.17%	Hebrew Israelites (Black group)	4.92%
white privilege	8.28%	essential workers	4.70%
how to build resilience	7.83%	hypocrites	4.70%
black girls rock	7.61%	masks	4.70%
artvsartist	7.38%	china lied	4.47%
stay home save lives	7.38%	minimum wage	4.47%
hand wara	7.16%	real relief now	4.47%
nurses	7.16%	star wars	4.47%
caucus	6.94%	Wisconsin	4.47%
families	6.71%	women matter	4.47%
stocks	6.49%	businesses	4.25%
humors	6.26%	disparities	4.25%
schooling	6.26%	end times	4.25%

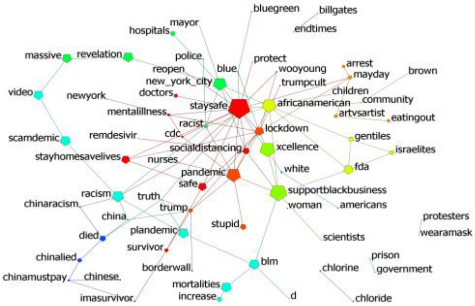

Figure 1. Semantic visualization of topic modeling on COVID-19 Tweets authored by African Americans

Fifteen topics (four themes) were identified in 1,763 publicly available COVID-19 related Tweets within African American Twitter communities (Newman modularity 0.74). Four main themes include: 1) Symptom and Transmission Patterns- cure, case, patient, death; 2) Treatments, Prophylaxis and Cures- data, CDC, mortality, rate; 3) Effectiveness/Impact of Interventions by Health Authorities and Other Institutions– day of lockdown, quarantine, covid safe, crisis, impact, free food; 4) Fear –protect, help, new, African American community, black, isolation, quarantinelife, school, foodshortage.

4. Discussion

We detected fifteen topics in publicly available Tweets by African Americans applying topic and further categorized those topics into four main themes including symptoms/transmission patters, treatment, implementation of interventions, and fear. Social media has become important in the global discussion of the science behind the pandemic. Understanding the spread of the virus is critical because of its deadly nature. Yet, several hundred new studies arise daily. The global scientific community is cooperating in ways never seen before in order to ensure vaccine and treatment development are guided by the most accurate findings [3]. In a study comparing COVID-19 knowledge by misinformation source found that social media respondents were unconvinced handwashing or social distancing was valuable. In fact, tremendous distrust from the CDC emerged as social network respondents believed the COVID-19 threat was a politically motivated hoax from the CDC, the virus is man made in China and Vitamin C is a secret cure [4,5]. Inaccurate information spread on social networks generally includes the impact from COVID-19 is not real including the death toll is a lie. Medical professionals are bombarded dispelling myths during care visits which is beginning to take a toll on their well-being [6]. Fear is evidence by people not seeking care when COVID-19 symptoms exist, the drinking of bleach indicating we are not doing a good job in protecting the public from inaccurate information. Yet, in the recent weeks social networks have taken a variety of steps to dispel myths including the addition of a portal to vet information from public officials and banning conspiracy theory content, which drives fear. Some public figures with popular social media platforms have been observed to use fear and chaos to drive their agendas and recruit people for their causes. They often spread distrust of the government and the information in the news about the virus, treatments and death. Follower with inaccurate information are often times confused leading to adverse effects including the ignoring of stay-at-home orders. The US media system supports disinformation, racism and hate as it drives up the ratings and advertising revenue, making dispelling such myths even more difficult. In March, however, Twitter set out to ban inaccurate information and to remove such influencers like Fox news hosts as it listened to civil rights organizations to tack these actions [6,7]. The generalizability of this study is limited by using the control, which did not exclude Tweets written by African American individuals. A few recent classifiers have been developed to detect demographic information in Twitter; however, they were either unavailable or in a form that is not easily applicable. As a result, some important topics uniquely discussed by African American Twitter communities may have been excluded.

5. Conclusions

Our topic modeling detected information, rumors, positive sentiments, cohesively encouraging online discussions (e.g., Black strong 27.1%, growing up Blacks 22.8%, support Black business 17.0%, how to build resilience 7.8%), and COVID -19 prevention behaviors (e.g., masks 4.7%, encouraging social distancing 9.4%) as unique culture within African American Twitter communities. This new knowledge adds insights regarding future virtual based intervention and social media based health disparity research for COVID-19.

Acknowledgments: U.S. federal grant R01AG060929.

References

[1] Yancy CW. COVID-19 and African Americans. JAMA 2020.

[2] Blei DM. Technical perspective: Expressive probabilistic models and scalable method of moments. Communications of the ACM 2018;61: 84-84.

[3] Boyle A. Studies of coronavirus evolution stir up a controversy for scientists on social media. geekwire.com; 2020.

[4] Pennycook G, McPhetres J, et al. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention; 2020.

[5] Singh L, Bansal S, et al. A first look at COVID-19 information and misinformation sharing on Twitter. arXiv preprint arXiv:2003.13907; 2020.

[6] Ferrara E. # COVID-19 on Twitter: Bots, Conspiracies, and Social Media Activism, arXiv preprint arXiv:2004.09531; 2020.

[7] Pakpour A, Griffiths M. The fear of CoVId-19 and its role in preventive behaviors. Journal of Concurrent Disorders 2020.