Using artificial intelligence to develop a lexicon-based African American Tweet detection algorithm to inform culturally sensitive Twitter based social support intervention for African American dementia caregivers

Broadwell, P., PhD,[1] Odlum, M., EdD,[2] Asim, H., BS,[3] Deng, N., BA,[3] Davis, N., PhD,[4] Alcantara, C., PhD,[5] Mittelman, M.S., DrPH[6], Yoon, S., PhD[2,7]
[1]Center for Interdisciplinary Digital Research, Stanford University, Stanford, CA, [2]Columbia University Irving Medical Center, New York, NY,
[3]School of Professional Studies, Columbia University, New York, NY, [4]School of Nursing, Clemson University, Clemson, SC,
[5]School of Social Work, Columbia University, New York, NY, [6]Department of Psychiatry, Grossman School of Medicine, New York University, NY
[7]Columbia University Data Science Institute, New York, NY

Background and Aims

The prevalence of dementia is higher for African Americans than Whites.[1] Although deep learning and other statistical techniques have been widely applied to infer demographic information on Twitter, those demographic detection algorithms tend to be unavailable to open science communities and/or require access to account details that could compromise individuals' privacy.[2,3] The purpose of this study is to develop a lexicon-based African American Tweet detection algorithm to inform culturally sensitive Twitter based social support intervention for African American dementia caregivers.

Methods

- For our Tweet corpora, we extracted 3,291,101 Tweets using hashtags associated with African American-related discourse (#BlackTwitter, #BlackLivesMatter, #StayWoke) and 1,382,441 Tweets from a control set (general or no hashtags) from September 1, 2019 to December 31, 2019 using the Twitter API.
- For our literature corpora, we extracted 14,692 poems and prose writings by African American authors and 66,083 items authored by others as a control, including poems, plays, short stories, novels and essays, using a cloud-based machine learning platform (Amazon SageMaker) via ProQuest TDM Studio.
- Lastly, we combined statistics from log likelihood and Fisher's exact tests as well as feature analysis of a batch-trained Naive Bayes classifier to select lexicons of terms most strongly associated with the target or control Tweets.

Data Processing

A total of 803,495 Tweets (24.41%) associated with African American-related discourse and 369,348 Tweets (26.71%) in the control group were identified as unique and non-bot generated Tweets.[4]

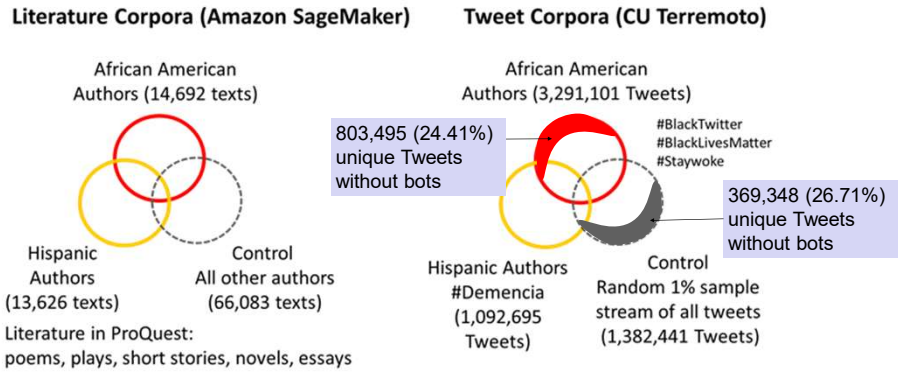

Figure 1. Volume of texts used to create African American lexicon and African American discourse-associated Tweet detection algorithm after applying a Twitter-bot detection process

Results and Discussion

- We found that a lexicon composed of unigrams was more effective at differentiating Tweets from held-out test samples of the two groups than lexicons composed of n-grams of various lengths. This is likely due to the relatively terse nature of Tweets.
- Size of the lexicon: African American lexicon contains 1,735 unigrams and control contains 2,267 unigrams (Table 1).

Table 1. Statistics of log likelihood, Fisher's exact tests, and feature analysis of a batch-trained Naive Bayes classifier on sample terms in African American lexicon

term	aa_freq	all_freq	log_ratio	log_p	fisher_ratio	fisher_p	bayes_nll	doc_freq	log_inv_freq
black	28873	52819	1.85	0.00	2.89	0.00	-6.22	29880	1.71
white	22301	51368	1.47	0.00	1.84	0.00	-6.50	31968	1.64
old	20267	63743	1.08	0.00	1.12	0.00	-6.69	40211	1.41
woman	18606	40940	1.54	0.00	2.00	0.00	-6.59	24799	1.90
hair	7693	23581	1.11	0.00	1.16	0.00	-7.37	18011	2.22
hell	3786	10955	1.17	0.00	1.26	0.00	-7.88	8891	2.92
hurt	2756	7413	1.26	0.00	1.42	0.00	-8.11	6175	3.29
Jesus	1687	4298	1.33	0.00	1.55	0.00	-8.45	2973	4.02
Harlem	1586	1778	3.03	0.00	19.77	0.00	-8.33	1116	5.00
hoped	1140	2658	1.46	0.00	1.80	0.00	-8.80	2517	4.18
cops	796	1840	1.47	0.00	1.82	0.00	-9.04	1284	4.86
afro	656	757	2.94	0.00	15.54	0.00	-9.10	566	5.68

Conclusion

Our first version of a lexicon-based African American Tweet detection algorithm developed using literature and Tweet texts will be useful to inform culturally sensitive Twitter-based social support intervention for African American dementia caregivers.

Acknowledgments

Using Twitter to Enhance the Social Support of Hispanic and Black Dementia Caregivers (Tweet-S2) RO1AG060929

References

1. Alzheimer's Association. Alzheimer's disease facts and figures. Alzheimer's & Dementia: The Journal of the Alzheimer's Association. 2019.
2. Vijayaraghavan P, Vosoughi S, Roy D. Twitter demographic classification using deep multi-modal multi-task learning. 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) 2017 Jul (pp. 478-483).
3. Sloan L, Morgan J, Burnap P, Williams M. Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. PloS one. 2015 Mar 2;10(3):e0115545.
4. Subrahmanian VS, Azaria A, Durst S, Kagan V, Galstyan A, Lerman K, Zhu L, Ferrara E, Flammini A, Menczer F. The DARPA Twitter bot challenge. Computer. 2016 Jun 13;49(6):38-46

COLUMBIA UNIVERSITY DATA SCIENCE INSTITUTE