

**Supplementing materials for iConference publication titled “Categorization and Comparison of Influential Twitter Users and Sources Referenced in Tweets for Two Health-Related Topics”**

In the method section, we described applying our lexicons to the data, for space restrictions in the paper, we are attaching the method section here again with references to the lexicon below.

## 1 Method

To identify the type of each URL, we code the URLs based on a domain lexicon created to classify expanded URL domains into 12 different categories: *News, Videos, Fake News, Social Media, Blog, Commercial, Government, Scientific, Educational institutions, Health Magazines, National or state professional medical societies and associations, Health Insurance*. URLs were pulled straight from user tweets and were classified based on mutually exclusive generated URL domains. Our lexicon is constructed by assimilating publicly available categorizations of web domains into the aforementioned categories [1] (see Table 1 below). To identify influential users, we used the pre-defined Klout Score obtained from Crimson Hexagon through Klout API, which is a measure of a user’s overall influence, it was shown previously as a valid measure of influence [2]. Based on the Klout Score, the top 25th percentile of users were considered to be influential [3]. Within the top 25th percentile, 150 tweets were randomly chosen for each health topic. Finally, a codebook was developed by incorporating past research on classifying social media users [4] (see table 2 below). This was used to categorize influential users into two main categories: individuals and organizations, as well as several sub-categories (as seen in Figure 2 below). Using the codebook, two coders were trained to review and assign each of the Twitter profiles to one of the categories identified above. The types of users were determined through each user’s personal user profile, mostly commonly with the information found in the user’s bio. The coders agreed on 252 out of 300 profiles (Cohen’s Kappa coefficient 81.5%).

**Table 1.** URL Categorization Code Book

#	Category	Description	Examples
1	News	An official broadcast or published report of current events, important information, etc.	Nytimes.com, Cnn.com, Theguardian.com
2	Blogs	A regularly updated website or web page, typically one run by an individual or small group, that is written in an informal or conversational style.	Blogspot.com, Wordpress.com, Tumblr.com
3	Scientific	Scholarly publications, research, articles or journals that contain credible scientific data/information.	Bmj.com, Elsevier.com, Nature.com
4	Federal, State or Local Government agencies	A government or state agency, often an appointed commission, is an organization in the machinery of government that is responsible for the oversight and administration of specific functions.	Nih.gov, Cdc.gov, medlineplus.gov Any .gov URL
5	Commercial content	Websites that are available to list and search for properties at no charge but may offer additional services on a fee basis.	Amazon.com, eBay.com
6	National or state professional medical societies and associations	An official national or state organization representing a particular group of medical professionals.	Mayoclinic.org, kaiserpermanente.org, Apa.org
7	Health magazine websites	Magazines that cover a variety of topics including physical fitness and well-being, nutrition, beauty, strength, bodybuilding, weight training, etc.	Health.com, Prevention.com, Self.com
8	Health Insurance	An insurance that covers the whole or a part of the risk of a person incurring medical expenses.	molinahealthcare.com, assuranthealth.com, unicare.com
9	Videos	Video sharing services where users can watch, share, upload their own videos, etc.	YouTube.com, Vimeo.com
10	Social Media	Websites that enable users to create and share content or to participate in social networking.	Facebook.com, Twitter.com Reddit.com, Pinterest.com
11	Educational institutions	Websites that represents a place where people of gain an education.	Berkeley.edu, Colorado.edu, Sdsu.edu, Any .edu URL

12	Fake News	A type of journalism or propaganda that consists of deliberate misinformation or hoaxes.	100percentfedup.com, 21stcenturywire.com
----	-----------	--	---

**Table 2.** User Categorization Code Book

	Dimension	Factor	Definition	Location
Individuals	Occupation	Health Professional	An individual who provides/is involved in health care services (doctor, physician, nurse, etc.)	Bio (Linked URL), Username
		Blogger	A personal log of thoughts published on a Web page or on the Twitter account itself and is maintained by one person.	Bio (Linked URL), Username
		Celebrity	An individual notable for their fame (musician, actor, entrepreneur, writer etc.)	Bio (Linked URL), Verification Status
		Health Writer	Writers who can be considered experts from taking part in published works such as books, papers, magazines, etc. Should write about health topics i.e. vaccines	Bio (Linked URL), Verification Status
		Company Representative	An individual authorized to act on behalf an organization (publicist, advertiser, promoter, etc.)	Bio (Linked URL)
		Health Activist	An individual who campaigns to bring about political or social change in the health field.	Bio (Linked URL)
		Lay Person	An individual who discusses health topics with no occupation involving the topic discussed. Does not have any other online content beside their tweets. i.e. do not have a blog	Bio (Linked URL)
		Other	Cannot be identified in a category within occupation.	
Organizations	Dimension	Factor	Definition	Location
	Media	Official	Well-known, credible news media sources such as news channels (BBC, CNN) and health magazines.	Username, Verification Status, Bio
		Unofficial	Private news outlets run by individuals/small groups of people. (Health Standards, personal blogs, etc.)	Username, Verification Status, Bio
		Blog	A regularly updated website or web page, run by a group of individuals, usually written in an informal or conversational style.	Username, Bio (Linked URL)
		Social Media Profile	An account that serves to provide tips and information related to health but is not linked to any other sources such as a blog. The only source of content comes from the account itself.	Username, Bio (Linked URL), Timeline
	Educational	Universities	Institution of Higher Education.	Username, Verification Status, Bio
		Hospitals	Health Care Institution.	Username, Verification Status, Bio
	Governmental	Government	Organization active in the field of global health research; dependent of the government	Username, Verification Status, Bio, URL domain name
	Nonprofits	Societies	A non-profit organization that specializes in improving healthcare.	Username, Verification Status, Bio
		Foundations	An organization designed to strive towards new research opportunities within a field in order to bring about improvement.	Username, Verification Status, Bio
		NGO's	Organization active in the field of global health research; independent of the government.	Username, Verification Status, Bio, URL domain name
	Health Corporations	Health-Care Companies	A profitable organization that provides health programs and services for all. (CVS Health, WebMD, etc.)	Username, Verification Status, Bio, URL domain name
	Other	Other	Cannot be identified in a category.	

## References

- [1] Addawood, A., Rezapour, R., Mishra, S., Schneider, J. and Diesner, J. Developing an information source lexicon
- [2] Rao, A., Spasojevic, N., Li, Z. and DSouza, T. 2015. Klout score: Measuring influence across multiple social networks. In *Proceedings of the Big Data (Big Data), 2015 IEEE International Conference on*. IEEE
- [3] Krauss, M.J., Sowles, S.J., Moreno, M., Zewdie, K., Grucza, R.A., Bierut, L.J. and Cavazos-Rehg, P.A. 2015. Peer reviewed: Hookah-related twitter chatter: A content analysis. *Preventing chronic disease*, 12
- [4] Liu, Y. 2016. Mining social media to understand consumers' health concerns and the public's opinion on controversial health topics