How can we improve social media analysis to better inform public health surveillance and practice? A Case Study in E-cigarette.

Annice Kim, PhD, MPH akim@rti.org
Center for Health Policy Science and Tobacco Research, RTI International, Berkeley CA

Introduction. Researchers are increasingly mining social media data to gain insights into public health issues. While social media data holds much promise, there are substantial challenges in acquiring, analyzing, and interpreting this data, and making this information relevant and accessible to public health practitioners. “There is currently a gap between what is possible with social media monitoring – which many studies have demonstrated successfully – with what is being done in practice (Paul MJ et al, 2016).” As a field, we need to move beyond content analysis to better understand who these data represent, in order to inform public health surveillance and evaluation efforts. In this paper, I describe several analyses from an ongoing NIH funded study to better understand who is sharing information about e-cigarettes on Twitter and how this data can be disseminated to public health researchers and practitioners.

E-cigarette Background. In recent years, e-cigarette use has increased rapidly in a largely unregulated marketplace and policy environment. With most users finding out about e-cigarettes through word-of-mouth and social media sources, platforms like Twitter have become an important data source for tracking the emerging e-cigarette phenomenon. Studies show that e-cigarette discussions on Twitter range from marketing and promoting e-cigarette products to policies regulating e-cigarette use, personal experiences with using e-cigarettes, and conversations about the risks and benefits associated with e-cigarette use. Monitoring e-cigarette social media discussions requires timely assessment of the content of posts but perhaps more importantly *who* is generating the content in order to inform regulatory and surveillance efforts. In 2016, the Food and Drug Administration (FDA) finalized a rule extending the agency’s authority to regulate e-cigarettes, which includes federal provisions requiring companies that sell e-cigarettes to include warning statements about nicotine on advertising/promotional materials, including content on digital/social media. To ensure that e-cigarette companies are complying with these advertising and labeling restrictions, FDA will need to identify and monitor websites and social media accounts originating from these companies. Furthermore, as public health researchers continue to use social media data to track and understand emerging e-cigarette perceptions and use behaviors, they will need to distinguish tweets from populations of interests (e.g. youth and young adults) vs content from other entities such as e-cigarette companies, marketers, or spammers who may be posting content for commercial purposes. Such information could also be useful in the development and targeting of social media campaigns to prevent e-cigarette use. Previous studies have used a range of techniques to identify Twitter accounts that are bots and promotional content that is posted by commercial entities, but no study has examined the diversity of user types engaging in e-cigarette conversations on Twitter, including e-cigarette/vaper enthusiasts, public health agencies, news organizations, and individuals.

Classification of Different Types of Users who Tweet about E-cigarettes. Using a supervised machine learning approach, we developed a method for classifying 5 different types of Twitter users who tweet about e-cigarettes: (1) individuals; (2) vaper enthusiasts; (3) informed agencies (public health agencies and news media); (4) marketers; and (5) spammers. A total of 4,897 users were manually classified according to the user type definitions Using a classification model that included metadata and tweeting behavior features, we were able to predict the five user types with relatively high accuracy (average F1 score = 83.5%). Accuracy varied by user type, with F1 scores highest for individuals (91.7%) and lowest for vaper enthusiasts (44.1%). Vaper enthusiasts were the most challenging user type to predict accurately and were commonly misclassified as marketers. Including additional tweet-derived features that capture tweeting behavior significantly improves the model performance beyond metadata features alone. Being able to segment e-cigarette twitter conversations by different user types of interest can better inform public health surveillance, education, and regulatory efforts.

Classification of Youth vs. Young Adults on Twitter. Public health organizations are increasingly using social media to disseminate messages about e-cigarettes to diverse audiences. Campaigns targeting youth and young adults actively use social media because it is an influential source of information in the lives of youth and young adults. Determining the extent to which the target audience was reached is critical to evaluating the impact of public health social media campaigns. Although audience demographic data is often reported in native analytic tools from social media platforms (e.g., Facebook Insights, Twitter Analytics), these tools have several limitations. First, the demographic information is not comprehensive across social media platforms and may be reported in categories that do not map to the target audience. Second, these analytic tools only provide demographic information about social media users who are actively following specific social media accounts (e.g., campaign Twitter handles) and not about users who may be actively discussing the campaign but not following these accounts. This limits researchers' ability to measure the true reach of their campaign efforts. Third, because these tools are proprietary, the methodological approach used to infer age or other demographic characteristics of social media users is unknown.

Using a supervised machine learning approach, we developed a method for classifying Twitter users into 3 age groups: (1) youth ages 13-17; (2) young adults ages 18-24; and (3) adults 25 or older. We were primarily interested in distinguishing youth and young adults given that these age groups are common targets for public health education campaigns and surveillance of risky health behaviors (e.g., tobacco, drug use, alcohol use, unprotected sex). We manually labeled 3184 Twitter users into one of the 3 age groups based on birthday announcement Tweets. The performance for our best model (74% precision, 74% recall, and 74% F1 score) was comparable with other three-class models predicting demographic characteristics. Our results suggest that models performed best with both linguistic and metadata features. By building age prediction models specifically for youth and young adult age groups that are at risk for negative health behaviors, our results can help inform better targeting of public health surveillance and education efforts on social media, as well as more target analyses of social media conversations among age groups of interest.

Dissemination. The presentation will end with a discussion on building social media analytic dashboards and the challenges of developing scalable systems and disseminating real-time information to public health audiences.