OJPHI

ISDS 2018 Conference Abstracts

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

# Leveraging Discussions on Reddit for Disease Surveillance

**Albert Park* and Mike Conway**

Biomedical Informatics, University of Utah, Salt Lake City, UT, USA

## Objective

We aim to explore how to effectively leverage social media for vaping electronic cigarette (e-cigarette) surveillance. This study examines how members of a social media platform called Reddit utilize topically-oriented sub-communities for e-cigarette discussions.

## Introduction

In recent years, individuals have been using social network sites like Facebook, Twitter, and Reddit to discuss health-related topics. These social media platforms consequently became new avenues for research and applications for researchers, for instance disease surveillance. Reddit, in particular, can potentially provide more in-depth contextual insights compared to Twitter, and Reddit members discuss potentially more diverse topics than Facebook members. However, identifying relevant discussions remains a challenge in large datasets like Reddit. Thus, much previous research using Reddit data focused on selected few topically-oriented sub-communities. Although such approach allows for topically focused datasets, a large portion of related data can be missed. In this research, we examine all sub-communities in which members are discussing e-cigarettes in order to determine if investigating these other sub-communities could result in a better smoking surveillance system.

## Methods

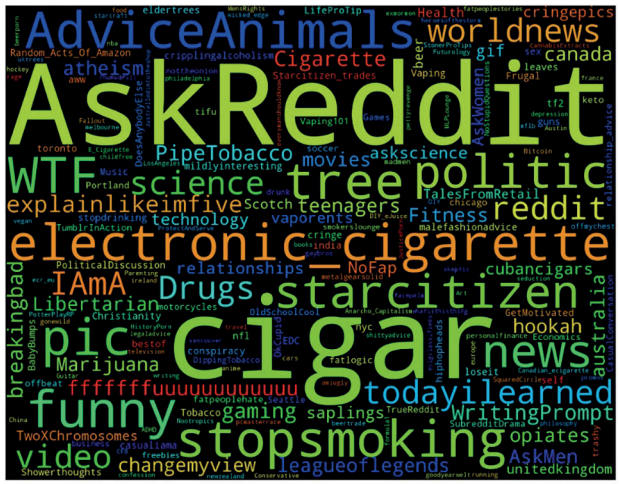
In this study, we use an archived Reddit dataset[1] that had been used in previous studies[2,3]. We first preprocessed the dataset, which included converting text to lower case and removing punctuation. Due to the size of the dataset (114,320,798 posts and 1,659,361,605 associated comments from 239,772 sub-communities), we identified 4 terms to extract posts or comments about e-cigarettes via a lexicon-based approach. The terms are 'e cig', 'elec cig', and 'electronic cig'. We included any partial matches in this process to cover a variation of e-cigarette terms. For example, a partial match of 'cig' can cover 'cig', 'cigs', 'cigarette', and 'cigarettes'. We presented the Wordcloud of the names and frequencies of sub-communities, in which members discussed e-cigarettes.

## Results

We extracted 354,587 posts/comments that were made by 176,252 unique member IDs from 6,039 unique sub-communities. There were 6 sub-communities with more than 8,000 e-cigarette posts. The sub-communities are 'AskReddit' (59,939) 'Cigars' (51,684) 'electronic_cigarette' (24,393), 'trees' (17,752), 'pics' (8,792), 'stopsmoking' (8,589). Other notable sub-communities are 'news' (5,010), 'politics' (4,662), 'worldnews' (3,785), 'science' (3,279), 'Drugs' (2,967), 'PipeTobacco' (2,099), 'Cigarettes' (1,401), 'teenagers' (1,016), 'AskMen' (918), 'Marijuana' (826), 'Fitness' (818), 'AskWomen' (698), 'cubancigars' (695), and 'vaporents' (608). Members were participating not only in sub-communities related to smoking and smoking cessation, but also in science, news, health, teenager, and Q&A sub-communities. The overview of the sub-communities that members participated to discuss e-cigarette are summarized in Figure 1.

## Conclusions

We present preliminary findings concerning the various sub-communities in which members had discussion on e-cigarettes in the popular social media platform Reddit. Our initial results suggest that Reddit members openly discuss electronic cigarette-related issues in many sub-communities that are unrelated to smoking. For the purpose of e-cigarettes surveillance, understanding the discussions in unrelated sub-communities, for example the subreddit 'teenagers', can provide opportunities to gain an in-depth perspective on the increased use of e-cigarettes by youth or non-smoker[4]. Moreover, high levels of activities in Q&A sub-communities like 'AskReddit', 'AskMen', and 'AskWomen' could indicate ineffective information dissemination regarding e-cigarettes[5], warranting further investigation. For the purpose of disease surveillance, we conclude that understanding the discussion in unrelated sub-communities has the potential to improve the practice of public health surveillance.




## Keywords

Data Mining; Surveillance System; Social Media; Electronic Cigarette; Smoking

## Acknowledgments

University of Utah's Institutional Review Board exempted the study procedure and data (IRB 00076188).

AP was funded by the National Library of Medicine of the National Institutes of Health under award number T15 LM007124.

MC was funded by the National Library of Medicine of the National Institutes of Health under award numbers R00LM011393 & K99LM011393.

## References

1. Reddit_Member. I have every publicly available Reddit comment for research. ~ 1.7 billion comments @ 250 GB compressed. Any interest in this? 2015. Archived at: http://www.webcitation.org/6kgAuNxDE

2. Park A, Conway M. Longitudinal Changes in Psychological States in Online Health Community Members: Understanding the Long-Term Effects of Participating in an Online Depression Community. J Med Internet Res. 2017 Mar 20;19(3):e71.
3. Park A, Conway M. Towards Tracking Opium Related Discussions in Social Media. Online J Public Health Inform
4. McMillen RC, Gottlieb MA, Shaefer RMW, Winickoff JP, Klein JD. Trends in Electronic Cigarette Use Among U.S. Adults: Use is Increasing in Both Smokers and Nonsmokers. Nicotine Tob Res. 2015 Oct;17(10):1195–202.
5. Park A, Zhu S-H, Conway M. The Readability of Electronic Cigarette Health Information and Advice: A Quantitative Analysis of Web-Based Information. JMIR public Heal Surveill. 2017 Jan 6;3(1):e1.

***Albert Park**
E-mail: alpark1216@gmail.com