What Do People Talk about When They Tweet about HPV?

Semantic and Sentiment Analyses of HPV and Cervical Cancer-Related Tweets

Background

Although there have been many national initiatives and campaigns to increase cervical cancer screening rates in the U.S., receipt of recommended Pap smear screening has stagnated, and uptake of human papilloma virus (HPV) vaccination remains low (Centers for Disease Control and Prevention, 2017). Much of the work on promoting cervical cancer preventative behaviors to date has focused on the importance of physician recommendations (Kepka, Ding, Hawkins, Warner, & Boucher, 2016), which clearly misses the significant number of women who do not have a regular source of heath care. Online social media that have transformed information sharing represent important new opportunities to examine women's beliefs and attitudes toward cervical cancer prevention behaviors on a large scale (Dunn, Leask, Zhou, Mandi, & Coiera, 2015). Analyses of social media content consists of spontaneous real-life health narratives and first-hand personal statements can make significant contributions to designing effective health campaigns (Yoo, Kim, & Lee, 2016). By applying semantic network analysis and sentiment analysis (Ruiz & Barnett, 2015) to a dataset of tweets related to cervical cancer and HPV vaccine, we aim to answer the following questions: (1) What topics were discussed and how did the topics change over time? (2) How did the topics differ regarding their sentiment? (3) How did topics and sentiments relate to types of tweet senders (e.g., government, organization, and celebrity, individual)?

Methods and Preliminary Results

We obtained an archived Twitter dataset containing a random 10% sample of all tweets from January 1, 2010, to December 31, 2014. Searching with keywords *pap smear*, *cervical* *cancer*, *HPV*, and *gardasil* yielded a dataset of 97,391 tweets. First, we conducted semantic network analyses on the whole dataset and five smaller datasets across five years. After removing stop words and stemming the texts, we used word co-occurrence (within five-word distance) to construct the semantic networks. Overall, the words follow a power law distribution with *HPV*, *cervical cancer*, *vaccine*, *pap smear*, and *woman* as the most frequently used words (Figure 1). Figure 2 shows the semantic networks for the whole dataset of tweets from 2010 to 2014. Overall, tweets fell into three topical categories: news, personal experiences, and general health information. Table 2 shows the modularity score and the top five keywords for each topic.

Analyses on each year's tweets revealed slight changes in word frequencies and topics, which were mostly driven by controversial topics happening in a particular year. For instance, in 2011, the word *Perry* was among the top 10 words, due to the fact that Rick Perry reversed his position on HPV vaccine mandate. In 2013, Gardasil became a salient word, because Merck implemented a voluntary recall of one lot of Gardasil.

We have finished the semantic network analyses and the next step is to conduct sentiment analyses to examine how topics differed with regard to their sentiments. We will apply machine learning procedure with a human-coded subsample as the training dataset. Analysis of variance will be used to test significant differences among sentiments. After that, we will examine the senders of the tweets. By using IBM Watson, we will classify the sources into four categories. New semantic networks will be constructed based on source classification to examine whether tweets from different sources exhibit different patterns in topics and sentiments.

Implications

Semantic network and sentiment analyses are useful tools to extract discussion topics on social media contents. Our findings suggest that as a social media platform, Twitter was not only used for distributing news, but also for sharing personal experiences. Public health campaigns can consider employing Twitter as a platform to reach people who are actively seeking and sharing health information.

References

Centers for Disease Control and Prevention (CDC). (2017, 8 24). HPV Vaccination Coverage Data. Retrieved from Centers for Disease Control and Prevention: https://www.cdc.gov/hpv/hcp/vacc-coverage.html

Dunn, A. D., Leask, J., Zhou, X., Mandi, K., & Coiera, E. (2015). Associations Between Exposure to and Expression of Negative Opinions About Human Papillomavirus Vaccines on Social Media: An Observational Study. Journal of Medical Internet Research, 17(6), 144.

Kepka, D., Ding, Q., Hawkins, A., Warner, E., & Boucher, K. (2016). Factors associated wit hearly adoption of the HPV vaccine in US male adolescents include Hispanic ethnicity and receiptof otehr vaccines. Preventive medicine reports, 4, 98-102.

Ruiz, J. B., & Barnett, G. A. (2015). Exploring the presentation of HPV information online: A semantic network analysisi of websites. Vaccine, 33(29), 3354-3359.

Yoo, S.-W., Kim, J., & Lee, Y. (2016). The Effect of Health Beliefs, Media Perceptions, and CommunicativeBehaviors on Health Behavioral Intention: An Integrated Health Campaign Model on Social Meida. Health Communication, 1-9.

Table 1: Topics of Tweets Related to HPV and Cervical Cancer.

Topics	Topic Modularity	Top 5 Keywords
News	48.80%	HPV, HPV vaccine, girl, vaccine, cancer
Personal experiences	31.02%	Pap smear, I, you, get
General health information	18.24%	cervical cancer, woman, symptom, screen, help

Figure 1: Top 20 Keywords' Frequency of Tweets Related to HPV and Cervical Cancer.

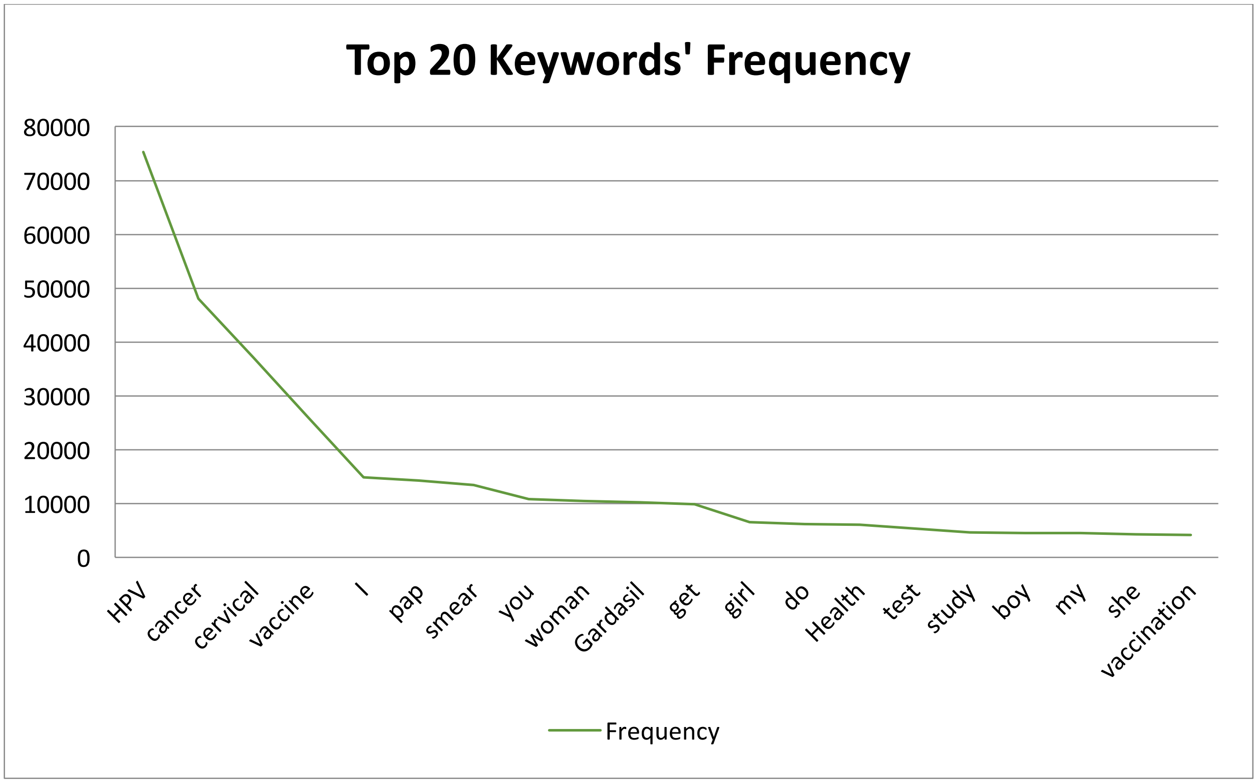
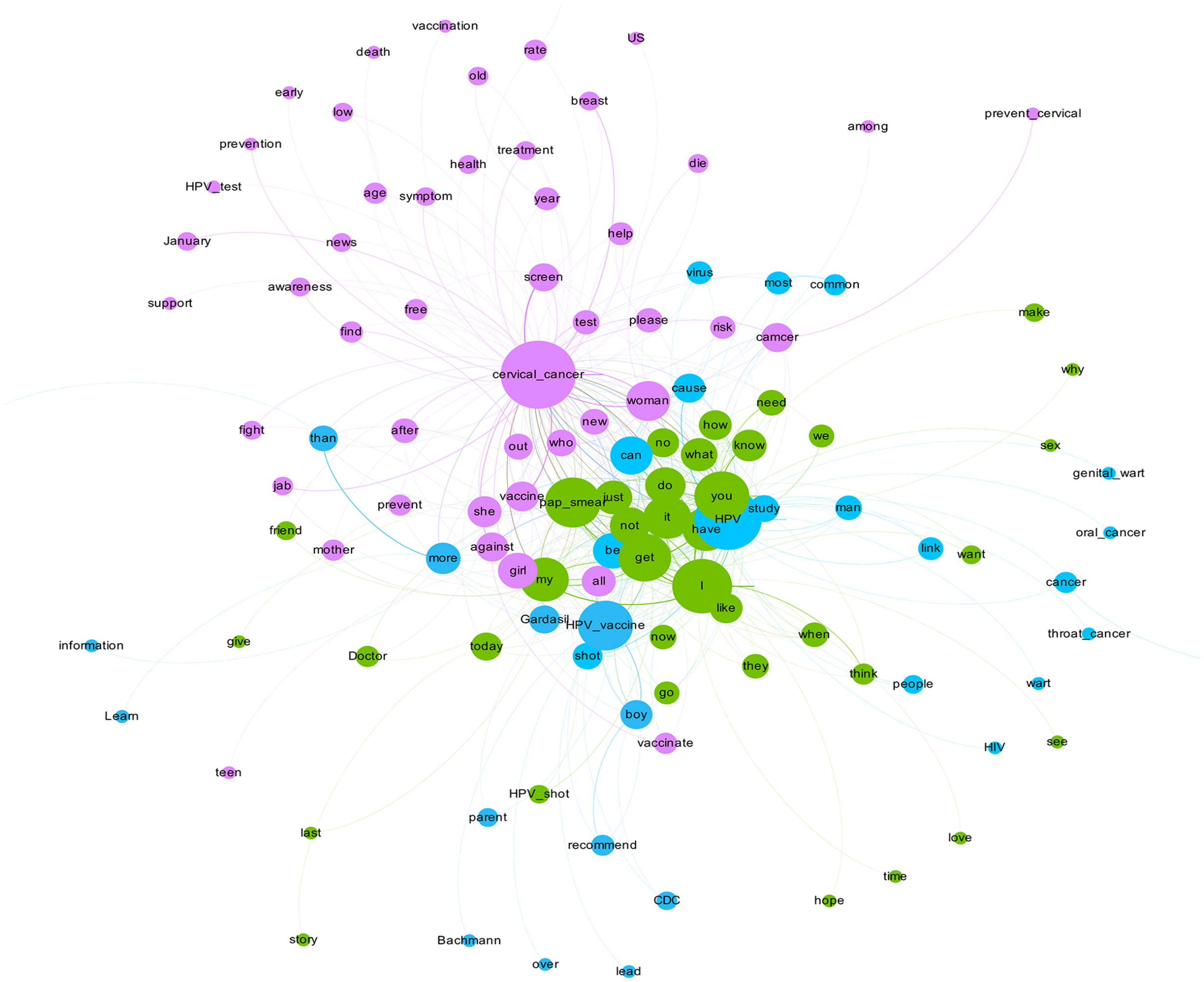


Figure 2: Semantic Networks of Tweets Related to HPV and Cervical Cancer.



Notes: The purple color represents the topic of news, the green color represents the topic of personal experience, and the blue color represents the topic of general health information.